Global Epidemic Analysis of Covid_19

Cathy Li
6/30/2020

1. Introduction

1.1 Background

At the close of 2019, a pneumonia of unknown cause was detected in the city of Wuhan in Hubei province, China. At the end of Jan 2020, the virus outbroken in Wuhan which named covid-19 by WHO. It seems this virus spread very fast, it outbroken in Italy and Spain in March then it swept US. As far as today, the totally confirmed case increased to more than two million. Many countries are falling into the crisis of lack of medical resources and hundreds people died every day. There is no specific medicine to treat this virus, ventilators are the key part to save lives, but its output is limited. Prevent virus is the most important thing for the whole world. Also the epidemic hit the world economic heavily, many industries are forced to stop business in order to prevent various spread, unanimous people lose their job and trapped into financial problems.

1.2 Problem

The whole world is concerned about epidemic, an intuitive epidemic map will help people better understand the global situation. Each country has different medical condition and capacity, this is the biggest element that influenced the country mortality rate. According to the big difference of infection rate by country, anti-epidemic measures definitely played a crucial role. So we need to use this data to decide what's the best useful way to prevent and control the spread of this virous and which country's medical system is worth for others to learn from.

2. Data acquisition

2.1 Data Resources

Recently, covid 19 is the most concerned issue around the world, WHO updates relevant data everyday and Johns Hopkins have designed a wonderful website with various of epidemic map and chart based on countries and areas. I got data from Enigma include the number of confirmed cases, deaths, recoveries by location and global, it also include geographic locations which could be used to draw maps, these csv files will be update daily in github, so I can retrieve the data easily thanks to theirs wonderful jobs. As for the population of countries and regions I use, these data are retrieved from Wiki.

2.2 Data Usage

There are three point I want to figure out, the first and also the most important part is Global Epidemic Map. As for this part, I need to get the number of confirmed cases by country and specific state or city and theirs longitude and latitude in order to generate map.

Sample Feature Selection

Global Confirmed Cases Trend			
Global Confirmed Cases	Date From	…	Date to
Total Number	1/22/2020	…	Current

Confirmed Cases Trend by Country/Region		
Country/Region	Daily Confirmed Case	Date From-To
China	***	1/22/2020 - Current
Japan	***	1/22/2020 - Current
Italy	***	1/22/2020 - Current
---	---	---

The secondary aspect I want to do research is Diagnosis Rate by Country and the most serious state in each country. I also need the data that used in before step and population of these countries and areas are necessary. The conclusion based on analysis will be helpful to provide some recommendations about what measures are effective to prevent virus.

Sample Feature Selection

Diagnosis Rate by Country/Region			
Country/Region	Total Confirmed Number	Population	Diagnosis Rate
China	***	***	***
Japan	***	***	***
Italy	***	***	***
---	---	---	---

At the last, Comprehensive Medical Level and Ability by Country is the topic that I'll seeking the conclusion by data. This analysis need to combine many elements, for example countries population, the percentage of confirmed case by countries population, the mortality rate by countries and by confirmed cases. Those data are key partial that could get some conclusions.

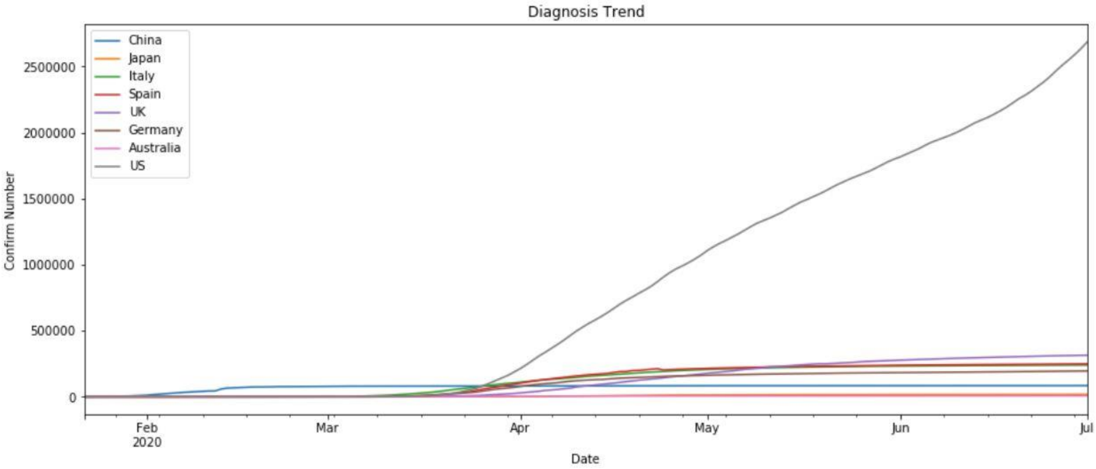
Sample Feature Selection

Comprehensive Medical Level and Ability Research				
Country/Region	Diagnosis Rate	Mortality Rate By Country	Mortality Rate By Confirmed Case	Population
China	***	***	***	***
Japan	***	***	***	***
Italy	***	***	***	***
---	---	---	---	---

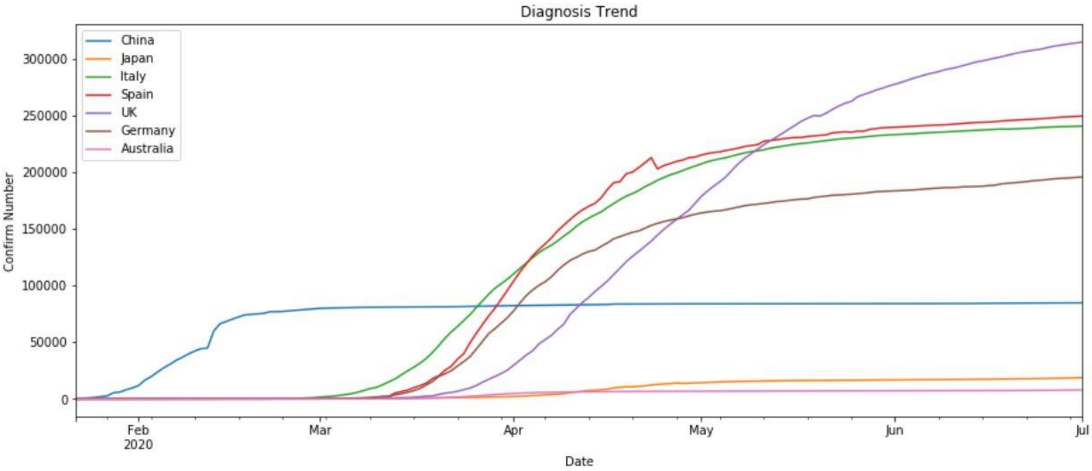
3. Exploratory Data Analysis

3.1 Trend of Pandemic

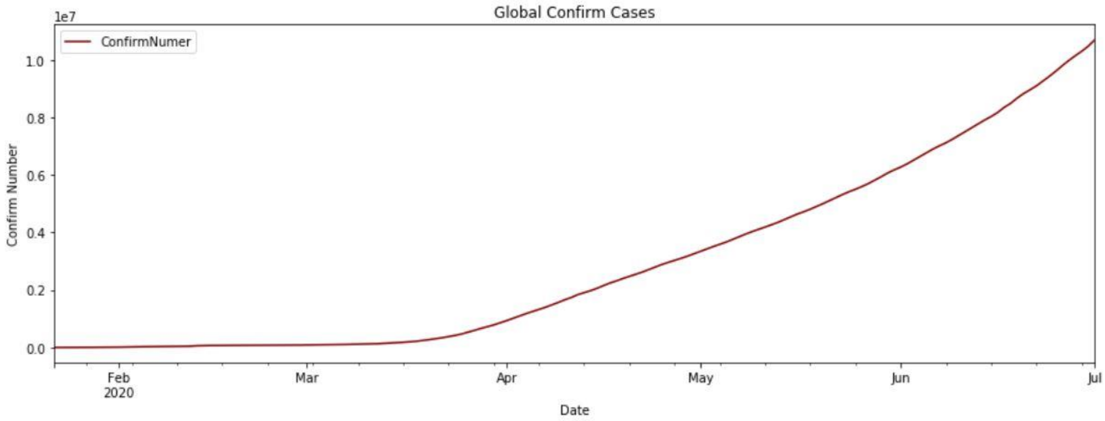
Global total confirmed number trend is the most directly way to understand pandemic development, I got the totally confirmed number through adding all confirmed numbers together by regions or countries and map the global development trend.



Eight Countries Diagnosis Trend

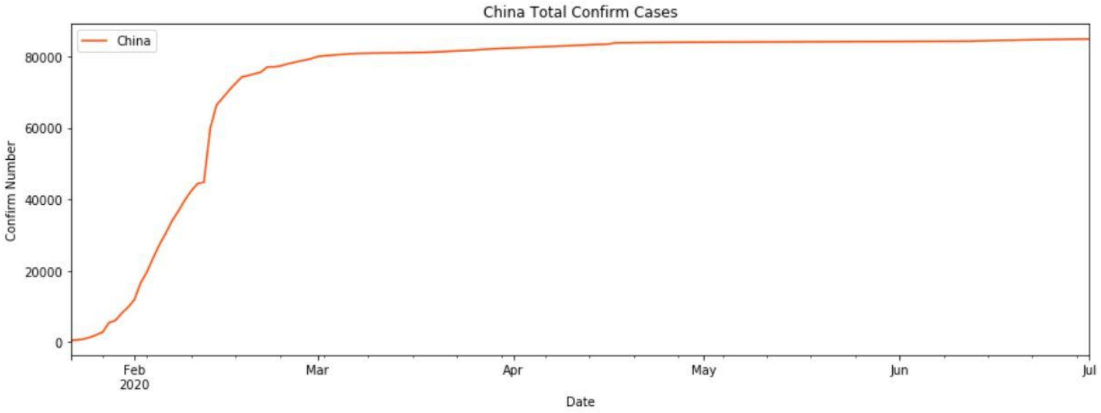


Seven Countries Diagnosis Trend

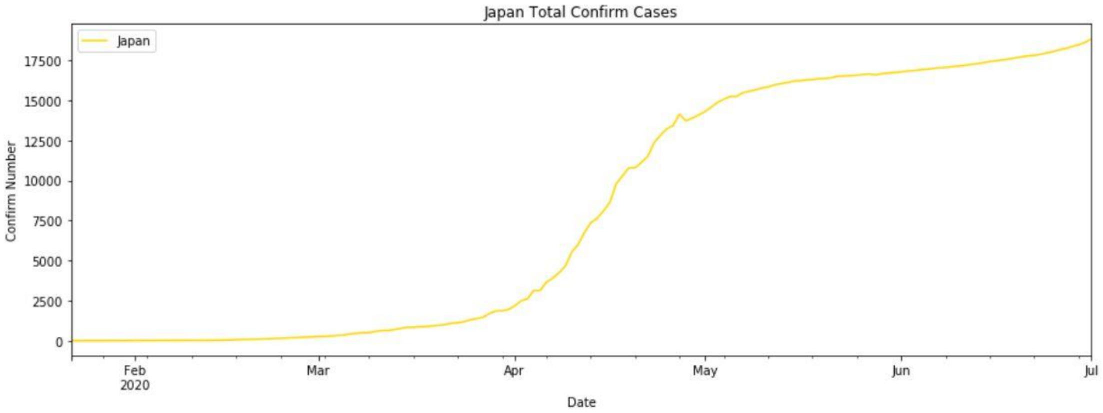


Global Totally Confirmed Cases Trend

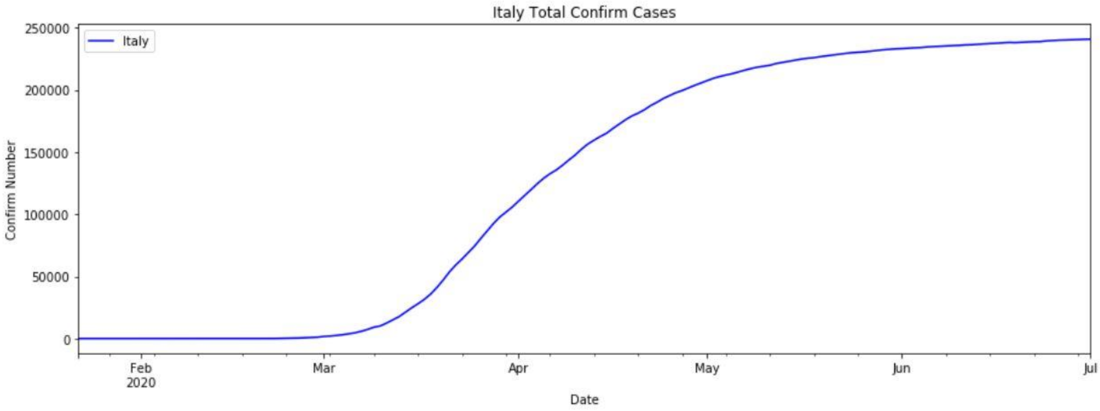
Also I have create same trend map for those representative countries in order to find out the what's the difference among those countries.



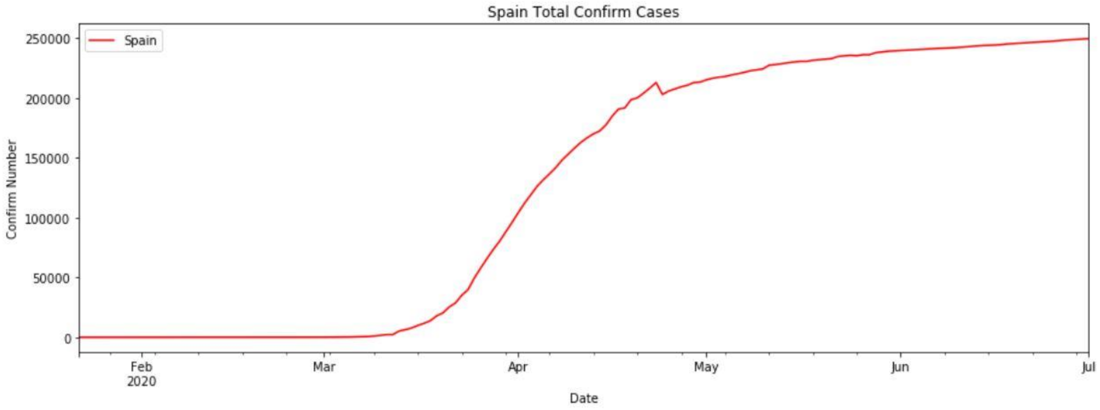
-China-



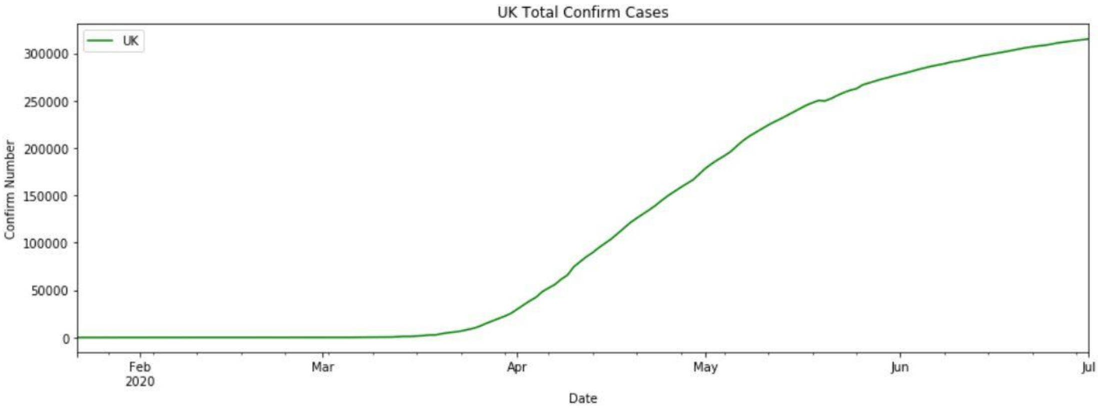
-Japan-



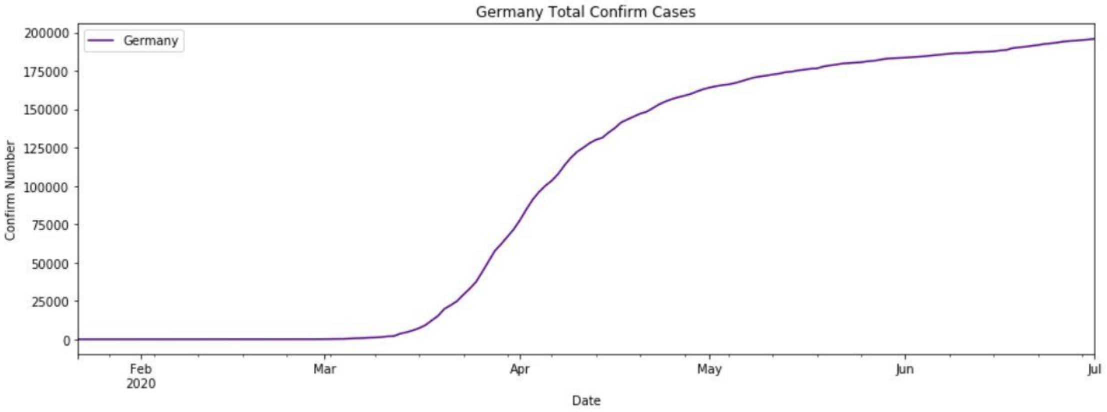
-Italy-



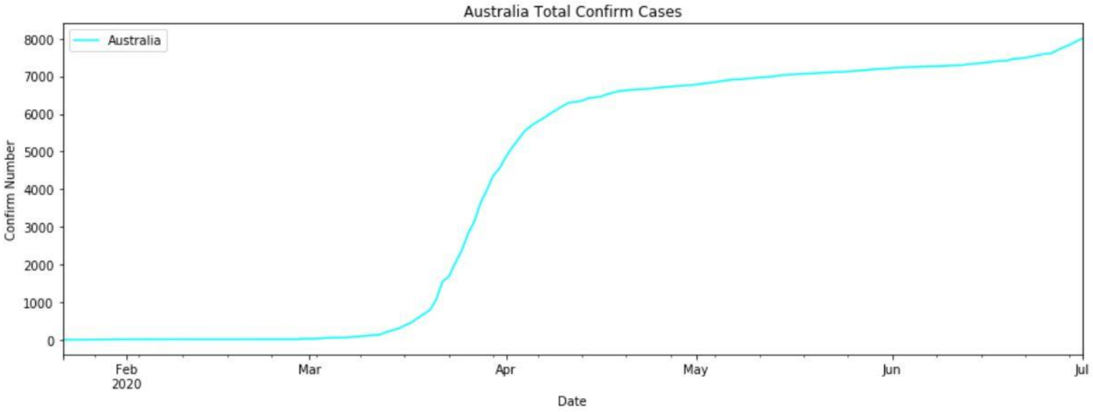
-Spain-



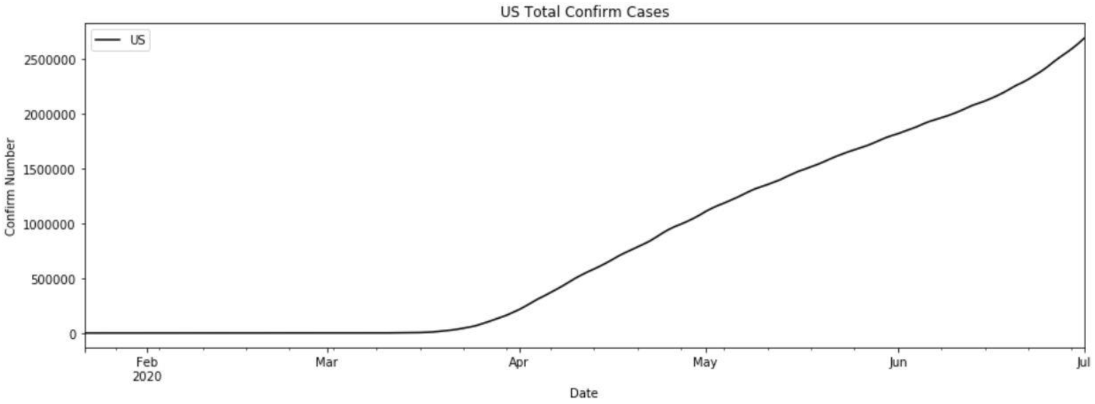
-UK-



-Germany-



-Australia-



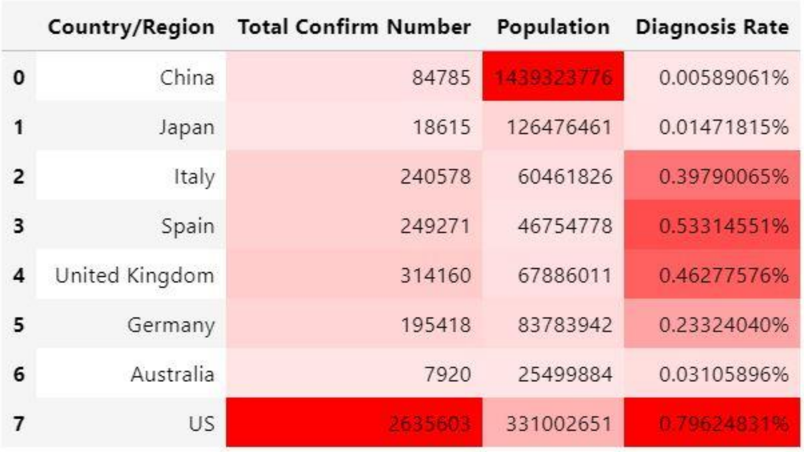
-US-

According to above pandemic trend maps, we can find out that all those countries trend are very similar except China and US, but China as the first outbreak country it has short time records before truly outbroken. Besides this difference, they all went through the same route and it seems US still in climbing stage, it’s hard to predict the turning point.

3.2 Diagnosis Rate Research

There must exist some cases that got this virus but not been confirmed, so the report maybe have some deviation because of this issue and here I just do my analysis based on the data that WHO published.

*Diagnosis Rate = Total Confirm Number/Population of Country

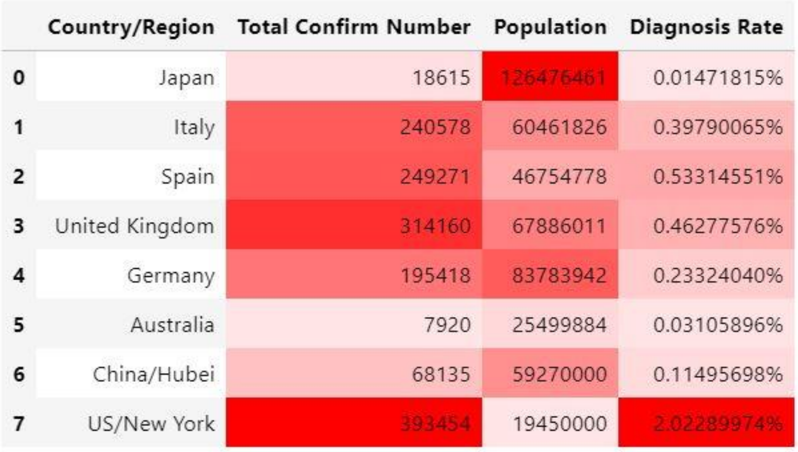


	Country/Region	Total Confirm Number	Population	Diagnosis Rate
0	China	84785	1439323776	0.00589061%
1	Japan	18615	126476461	0.01471815%
2	Italy	240578	60461826	0.39790065%
3	Spain	249271	46754778	0.53314551%
4	United Kingdom	314160	67886011	0.46277576%
5	Germany	195418	83783942	0.23324040%
6	Australia	7920	25499884	0.03105896%
7	US	2635603	331002651	0.79624831%

*Date from 22/1/2020 to 30/6/2020

As we all know that China, US and Australis, all of three has vast territory, so it's not a good way to compare them with other countries. Here I decide to pick up the most heavy state or province from China and US, then keep Australia as before because it's confirmed number are low.

Replace (China-Hubei Province; US-New York State)



	Country/Region	Total Confirm Number	Population	Diagnosis Rate
0	Japan	18615	126476461	0.01471815%
1	Italy	240578	60461826	0.39790065%
2	Spain	249271	46754778	0.53314551%
3	United Kingdom	314160	67886011	0.46277576%
4	Germany	195418	83783942	0.23324040%
5	Australia	7920	25499884	0.03105896%
6	China/Hubei	68135	59270000	0.11495698%
7	US/New York	393454	19450000	2.02289974%

*Date from 22/1/2020 to 30/6/2020

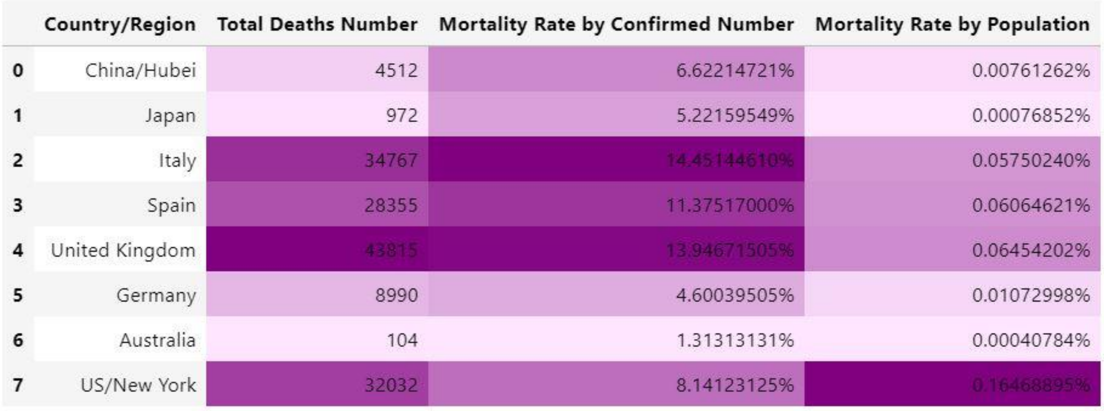
Here we can see, Japan has the lowest country diagnosis rate and New York has the highest region diagnosis rate. The second lowest country is Australia and not much difference with Japan. We could consider both countries diagnosis rate as level 1.

As for Europe countries, German has the lowest diagnosis rate, but Italy, Spain and UK all in similar aspect I'll consider them as level 2 countries. Hubei as the representative region of China, it should belong to level 2 refer to below data.
The last regions which should be in level 3 is New York, NY's diagnosis rate is much higher than other countries and regions.

3.3 Mortality Rate Analysis
Mortality Rate is a obvious way to assess medical level of a country or region, here I'll create a list of country mortality rate base on below countries and region

*Mortality Rate = Total Deaths Number/Total Confirme Number

*Mortality Rate = Total Deaths Number/Population of Country



	Country/Region	Total Deaths Number	Mortality Rate by Confirmed Number	Mortality Rate by Population
0	China/Hubei	4512	6.62214721%	0.00761262%
1	Japan	972	5.22159549%	0.00076852%
2	Italy	34767	14.45144610%	0.05750240%
3	Spain	28355	11.37517000%	0.06064621%
4	United Kingdom	43815	13.94671505%	0.06454202%
5	Germany	8990	4.60039505%	0.01072998%
6	Australia	104	1.31313131%	0.00040784%
7	US/New York	32032	8.14123125%	0.16468895%

*Date from 22/1/2020 to 30/6/2020

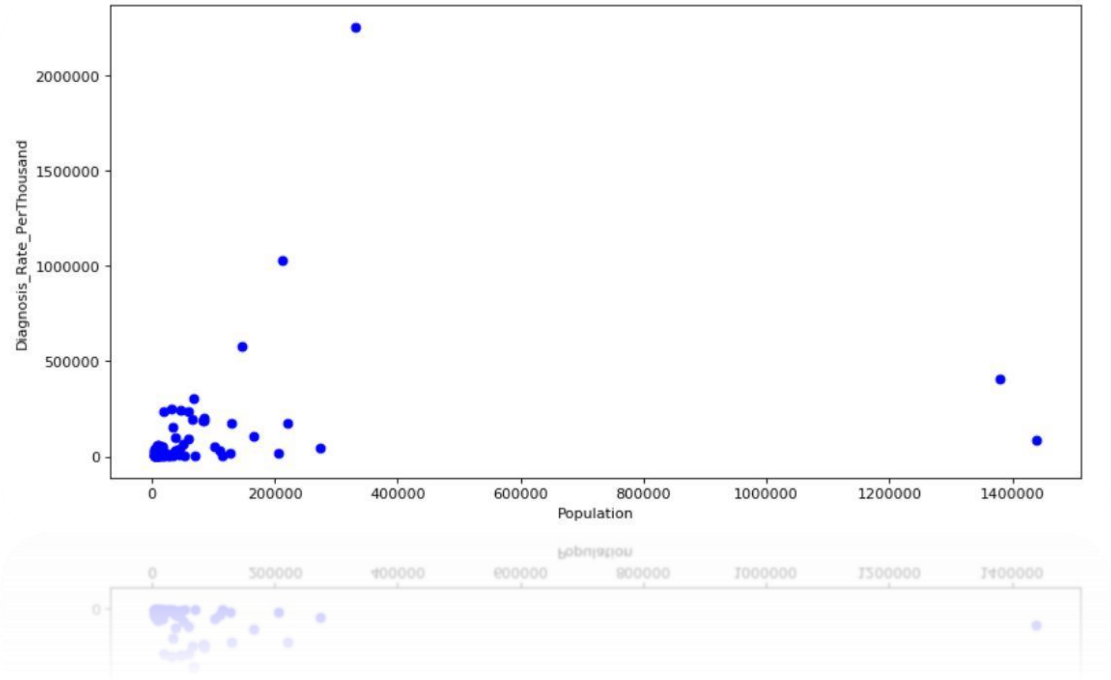
I'll divide below countries or regions to three ranks, according to mortality rate by confirm number. As for mortality rate by confirm number, the data of Australia and Germany both lower than 5%, so they belong to rank_1 without a doubt. The data of Hubei/China, Japan and New York/US all between 5% to 10%, all of them should be divide to rank_2. At last, Italy, Spain and UK, which data over 10% all in rank_3.

Comprehensive above data, Japan and Australia did the best job at controlling coronavirus spreading and also Australia has the lowest mortality rate either by confirm number or country population, that's proved it outstanding medical capability. As for medical capability, Germany and New York, US also in good performance even they has a big amount confirmed cases, it also shows their adequate medical resources.

4. Predictive Modeling

4.1 Build Model with countries confirm number and population

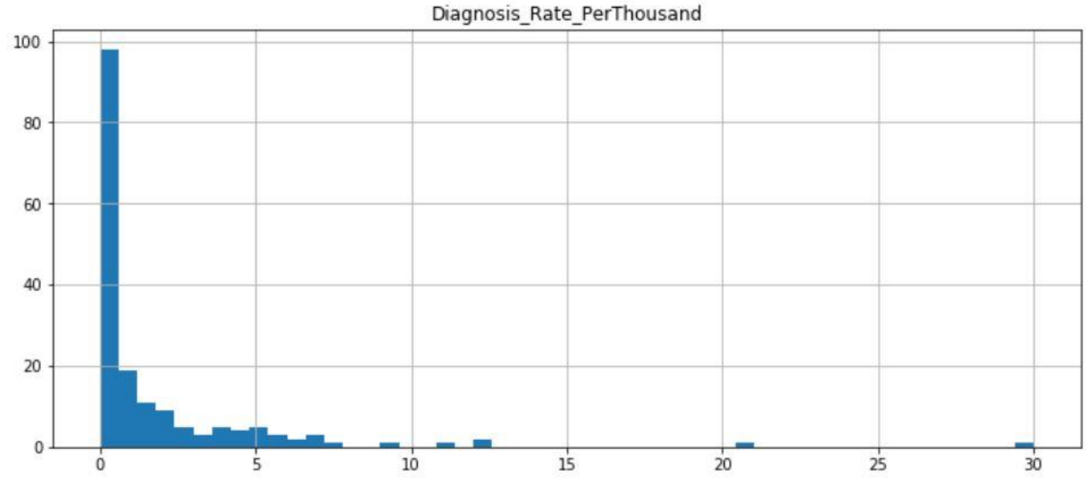
First step is processing data, here I chose to drop some too small simple which countries population less than 4 million or confirm number smaller than 1.5 thousands. Then plot scatter figure with each countries Diagnosis_Rate_PerThousand and Population.

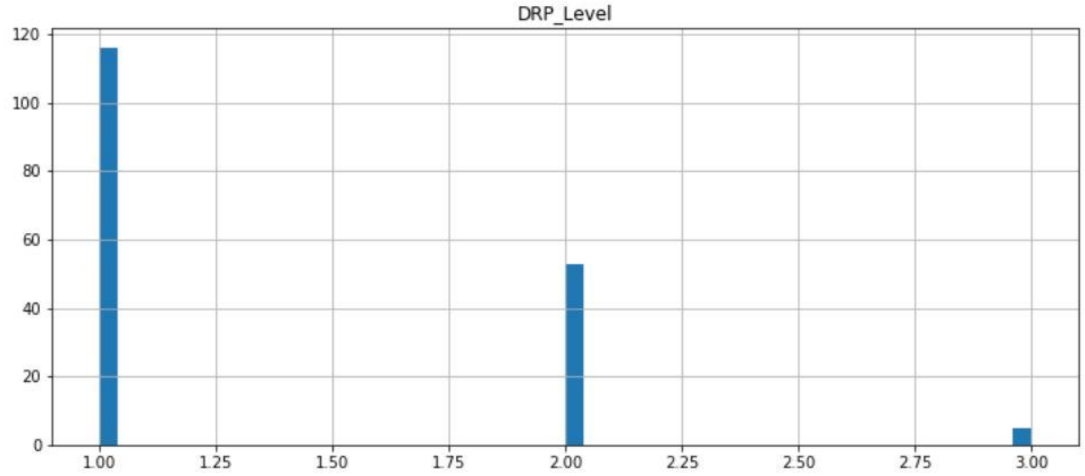


*Global Diagnosis Rate - Population

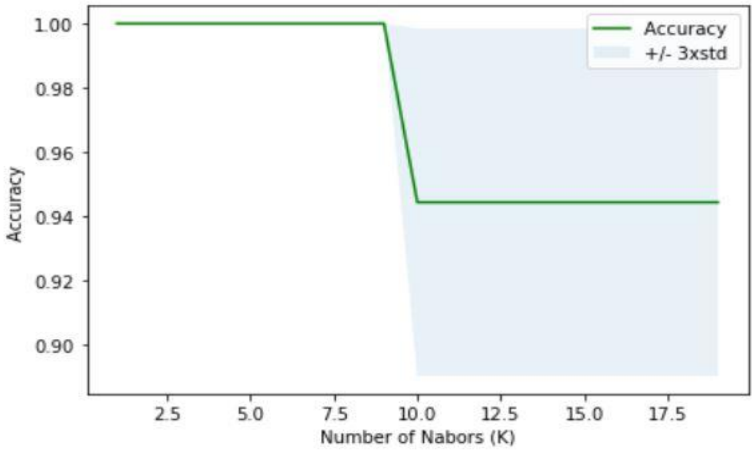
According to this scatter figure, the whole dataframe can't be fit to basic machine learning methods like linear regression, none linear regression or clustering. So, in order to modeling the data I'll import a new parameter to divide those date into several parts.

Refer to last part's diagnosis analysis, I'll mark the diagnosis rate as three level name DRP (Diagnosis Rate Per_Thousand) Level 1-3. After level sort process, here I got 116 countries/regions in DRP Level-1, 53 countries/regions in DRP Level-2 and 5 countries/regions in DRP Level-3, then I'll use K-Nearest Neighbors methods to build my predict model.
Before that, I'll build 2 hist map to show the important attributes data's distribution about DRP and DRP Level.





From above hist chart, the weight of infection rate decreases with the increase of diagnosis rate and therefore most countries or regions are in DRP Level 1, this sounds like a good news.



It’s obvious that the best K’s range are in 0 to 10 from above figure, here I’ll take the median k = 5. Then let’s calculate the KNN models accuracy while k is 5:

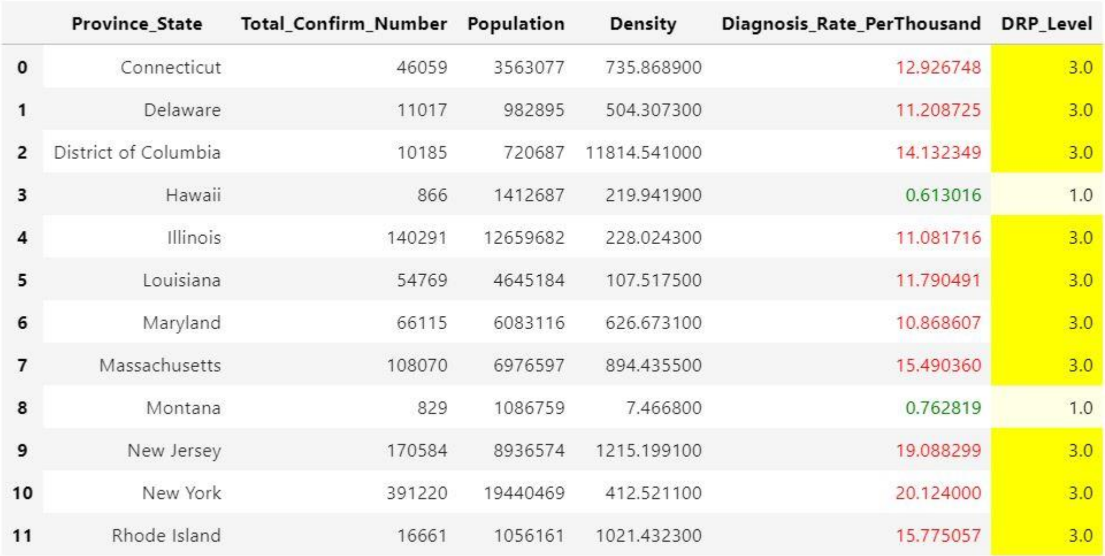
Train set Accuracy: 0.9615384615384616

Test set Accuracy: 1.0

Seems it’s really an ideal model, but because our test data are based on countries/regions, that’s really a limit for model building. So I cut down the test train data proportion in order to have more data to train the model, that maybe not a good way to resolve this issue.

4.2 Predict Severity of Outbreaks base on Country

First I will use this model to predict the most worst-outbreaks country’s situation, base on it’s states or provinces, to see if the results can fit the really tendency. So I use 52 states data as test train to predict theirs DRP_Level and got the results.

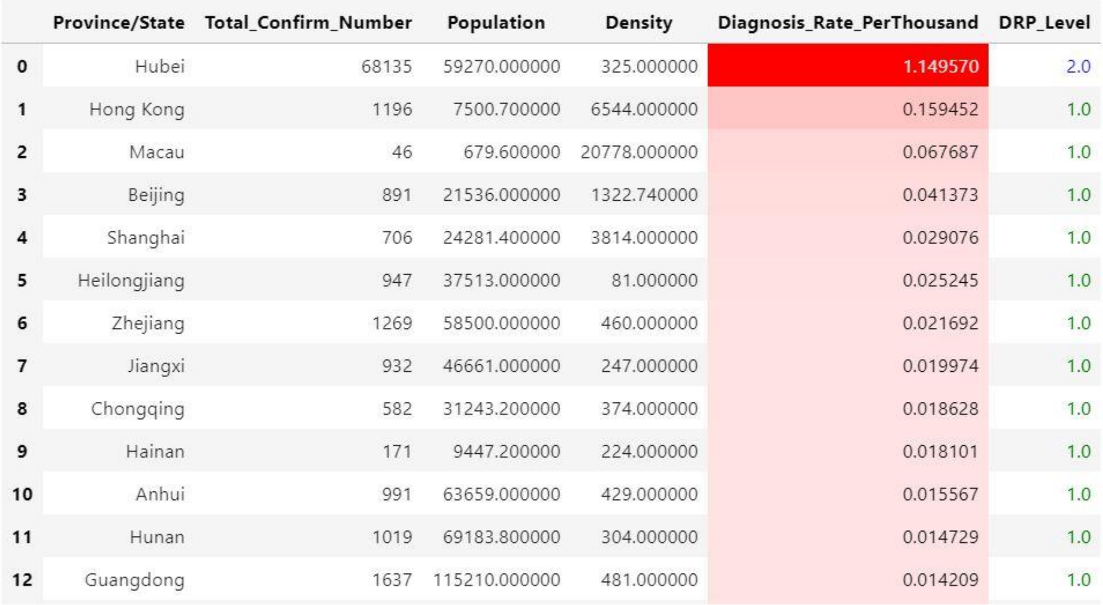


	Province_State	Total_Confirm_Number	Population	Density	Diagnosis_Rate_PerThousand	DRP_Level
0	Connecticut	46059	3563077	735.868900	12.926748	3.0
1	Delaware	11017	982895	504.307300	11.208725	3.0
2	District of Columbia	10185	720687	11814.541000	14.132349	3.0
3	Hawaii	866	1412687	219.941900	0.613016	1.0
4	Illinois	140291	12659682	228.024300	11.081716	3.0
5	Louisiana	54769	4645184	107.517500	11.790491	3.0
6	Maryland	66115	6083116	626.673100	10.868607	3.0
7	Massachusetts	108070	6976597	894.435500	15.490360	3.0
8	Montana	829	1086759	7.466800	0.762819	1.0
9	New Jersey	170584	8936574	1215.199100	19.088299	3.0
10	New York	391220	19440469	412.521100	20.124000	3.0
11	Rhode Island	16661	1056161	1021.432300	15.775057	3.0

*US Real Data

Here I only list the states that predict result not match it's really result as below.
As expected, only few data has deviation, now calculate it's accuracy as:
Test set Accuracy: 0.7692307692307693

Seems this is a not bad figure, then if we consider US as an extremely epidemic outbreak case, now maybe we can use this model to predict China's most worst epidemic trend if they aren't treat the crisis so serious.



	Province/State	Total_Confirm_Number	Population	Density	Diagnosis_Rate_PerThousand	DRP_Level
0	Hubei	68135	59270.000000	325.000000	1.149570	2.0
1	Hong Kong	1196	7500.700000	6544.000000	0.159452	1.0
2	Macau	46	679.600000	20778.000000	0.067687	1.0
3	Beijing	891	21536.000000	1322.740000	0.041373	1.0
4	Shanghai	706	24281.400000	3814.000000	0.029076	1.0
5	Heilongjiang	947	37513.000000	81.000000	0.025245	1.0
6	Zhejiang	1269	58500.000000	460.000000	0.021692	1.0
7	Jiangxi	932	46661.000000	247.000000	0.019974	1.0
8	Chongqing	582	31243.200000	374.000000	0.018628	1.0
9	Hainan	171	9447.200000	224.000000	0.018101	1.0
10	Anhui	991	63659.000000	429.000000	0.015567	1.0
11	Hunan	1019	69183.800000	304.000000	0.014729	1.0
12	Guangdong	1637	115210.000000	481.000000	0.014209	1.0

*China Real Data

According to predict data set, all province are in DRP Level 2 except Macau is belong to DRP Level 1. This is only my assumption, the development of epidemic is affected by many factors in fact, here I simplified model parameters in order to not beyond my knowledges.

So there is no need to calculate China's DRP Level predict data accuracy, it's obvious that accuracy value is very low. Only Hubei province meet the model predict value, I guess because Wuhan, the capital city of Hubei, is the first outbreaks area in the world. At the first period before the crisis been public, nearly nobody realized to do something to be safe like wear masks.

5. Discussion

As I said before, the covid-19 trend of each country or region is based on multiple attributes, like population, density, proportion of elderly population, national culture, national finance and production capacity, ect. During the data process and analysis, I have noticed a strange phenomenon that I thought the virus transmission speed is proportional to the population density but the truth is just the opposite. Although this two attributes not have absolute proportional relationship, from several typical countries, they have an inverse proportional relationship.

Asia is one of the most densely populated continents, compare to Asia it seems Europe and American are much more spacious, but theirs diagnosis rate are much higher than Asia countries like Japan, Korean and China. Why this happened? After a widely research, I found it happened partial because the culture are so different between eastern countries and western countries. As we all knew the covid-19 is spread through respiratory system, mask and social distancing are the most effective way to prevent virus spread. Compare to each other, eastern people more like to wear mask and wearing mask is even a fashion in some area like Korea and China, western people are more like to social, have a party or have a drink with colleagues after work. So, this culture gap lead to different road, we can't say which is right, it's a choice.

Another reason that Asia's diagnosis rate are lower than Europe and American is that several Asian countries had been through epidemic virus crisis in the last two decades, example SARS, MERS, both are belong to Coronavirus and cause respiratory system problems, so those people will treat covid-19 more seriously based on previous experience. The area that closed to Wuhan like Hong Kong, Macau, Taiwan, they stopped people entry permit from Wuhan or China mainland at the first time and require citizens to wear masks outside.

After above analysis, there is a special country that has the highest DR-D (DiagnosisRate/Density) in Europe, it's Sweden. Sweden is one of the countries with the lowest population density in Europe, but they didn't take any mandatory policy to pretend the virus and it also turns out, lock down cities, stop non-essential business, task some physical pretend measures are really works.

6. Conclusion

Maybe this is the biggest crisis in human history, it’s almost been 6 months from it started, hundreds of thousand people died because covid-19 and the world economy been hit hard, people losing their jobs everywhere.
Now, the most important thing is stop the virus spreading and develop vaccines, every country or region should learn something from other countries effective policy and methods. And It’s not a good time to dispute in-relevant things, saving life should be put on the first. According to those data, Herd Immunity is definitely unfeasible, we can’t afford the price at all.

At last, from this simple data analysis about global covid-19 epidemic, the obvious point is Wearing Mask is the simplest and most effective way that we can do to protect each other and help the world back to normal as soon as possible, please wear masks when necessary.