COVID-19 ANALYSIS REPORT

DS5010 – Intro to Programming for Data Science
2020 Fall

Yuan Ran
NUID 001569510
Instructor: Prof. Roi Yehoshua

Northeastern University

I. Abstract

The COVID-19 pandemic has been affecting the whole world negatively. It has led to dramatic loss of people's life over the world and disrupted people's normal life. In this report, we will focus on analyzing COVID-19 data to get a fundamental understanding of the situation worldwide and gain insights about the relationship between COVID-19 and climate. We will write a Python script and implement some useful libraries, like pandas, Numpy, Matplotlib, to do analysis. By importing Numpy and Pandas libraries, we will be able to load data from other sources to Jupyter Notebook, and transform it to a data frame, which is convenient for us to work on the dataset. Then we will use Matplotlib, which a magic plot tool, to plot different graphs to solve some questions about COVID-19. Finally, by importing climate json file, we will do some analysis to test the hypothesis that the spread of the virus is slowed down by warm weather.

II. Introduction

Facing the unprecedented challenge in the human history, scientists from different fields has been devoting themselves to defeat it. They have built a strong system to track and update COVID-19 global, which is convenient for people to check and be informed of the situation of COVID-19. Besides, they have even created the Exposure Notification System which can be installed on the phone. That helps people understand whether they've been exposed to someone who reports having COVID-19, and thus gives them health warnings and reminders.

The aim of this study is to help people understand the growth and cure of COVID-19 around the world, and thus give people a clearer and more direct picture of COVID-19. After analysis, we will answer questions like what is the total number of confirmed cases and number of deaths in each country in the last reported day, which countries present exponential growth in the number of cases and which countries are already leaving exponential growth. By plotting graphs, we can also get a clear picture that the number of deaths per 100 confirmed cases (observed case-fatality ratio) for the 20 most affected countries; After importing the world population json file and by computing the ratio between the total number of confirmed cases and the population size for each country, we can answer the question that what are the 10 countries with the highest number of confirmed COVID-19 cases per capita.

In the rest of the reports, we will talk about three essential parts which are composed of Data Acquisition, Data Analysis, and Conclusions. In the part of Data Acquisition, we will present a description of the used data sets, describe how the data was acquired and from where (with proper references), and describe any processing procedure used to prepare the data for analysis. In the part of Data Analysis, we will describe how to extract relevant information from the data sets and present the results for the different parts of the project and discuss them. In the conclusion, we will give our analysis results and share some insights about this study.

III. Data Acquisition

COVID-19 data is provided by John Hopkins University. The data provided contains a daily level information on the number of COVID-19 affected cases across the globe. The data is organized in a CSV file with the following columns:

- Date - observation date in yyyy/mm/dd. Country/Region - country or region.

- Province/State - province or state. Note that some countries do not have province or state and have empty values for this field.

- Confirmed - cumulative number of confirmed cases.
- Recovered - cumulative number of recovered cases.
- Deaths - cumulative number of deaths cases.

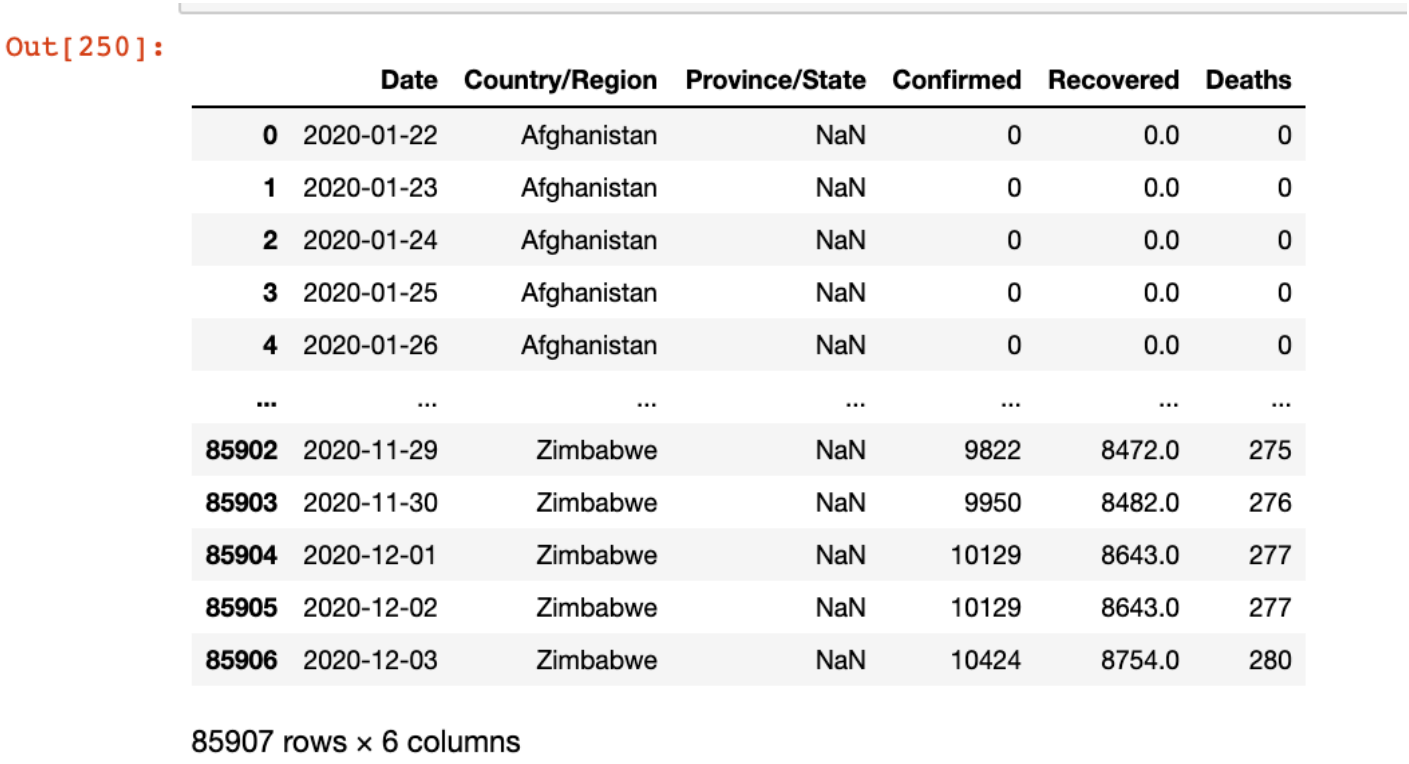
We Created a DataFrame from the CSV file to do analysis, and then merged the data for countries with multiple regions in order to provide a single time-series for each country and analyze data in a country level. Here is the reference to the data: https://raw.githubusercontent.com/datasets/covid-19/master/data/time-series-19-covid-combined.csv

Worldpopulation json file was used to join with COVID-19 data to put population data in each country into the created DataFrame and get a picture of ratio of affected cases in each country.

We also loaded climate json file which contains monthly climate date from over 100 stations around the world to test the hypothesis that the spread of the virus is slowed down by warm weather. We firstly normalized the json file by importing json module, and then merged climate data with COVID-19 data to do analysis.

IV. Data Analysis

To do the following analysis, we need to import several python libraries, like Numpy, Pandas, json, matplotlib. Firstly, we use pandas to read COVID-19 csv file and then transform it to a data frame. Here is the whole data set in the form of data format:



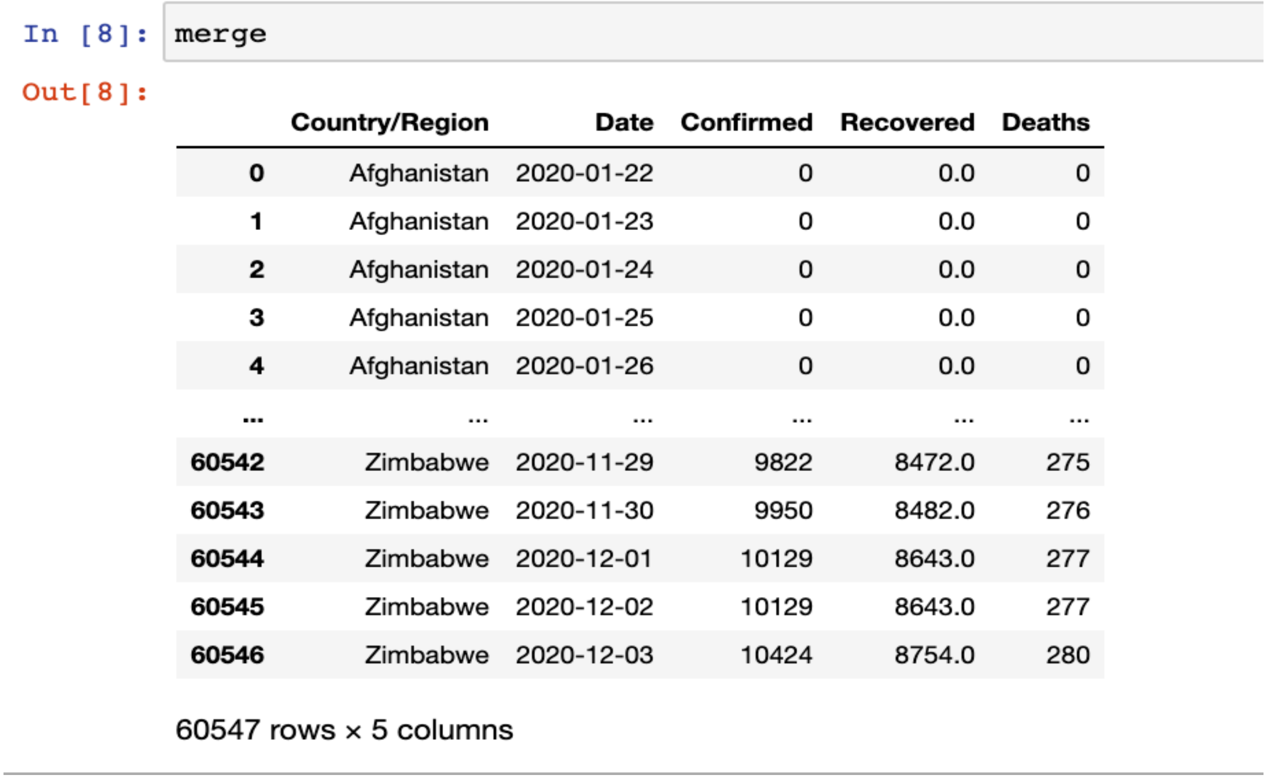
Out[250]:

	Date	Country/Region	Province/State	Confirmed	Recovered	Deaths
0	2020-01-22	Afghanistan	NaN	0	0.0	0
1	2020-01-23	Afghanistan	NaN	0	0.0	0
2	2020-01-24	Afghanistan	NaN	0	0.0	0
3	2020-01-25	Afghanistan	NaN	0	0.0	0
4	2020-01-26	Afghanistan	NaN	0	0.0	0
...	...	...	...	...	...	...
85902	2020-11-29	Zimbabwe	NaN	9822	8472.0	275
85903	2020-11-30	Zimbabwe	NaN	9950	8482.0	276
85904	2020-12-01	Zimbabwe	NaN	10129	8643.0	277
85905	2020-12-02	Zimbabwe	NaN	10129	8643.0	277
85906	2020-12-03	Zimbabwe	NaN	10424	8754.0	280

85907 rows × 6 columns

Figure 1

Then we clean the data by merging the data for countries with multiple regions in order to provide a single time-series for each country. To achieve this goal, we use the method of group by in pandas and sum up the columns of numbers. After implementing the method, we get the result like this:



```
In [8]: merge
Out[8]:
```

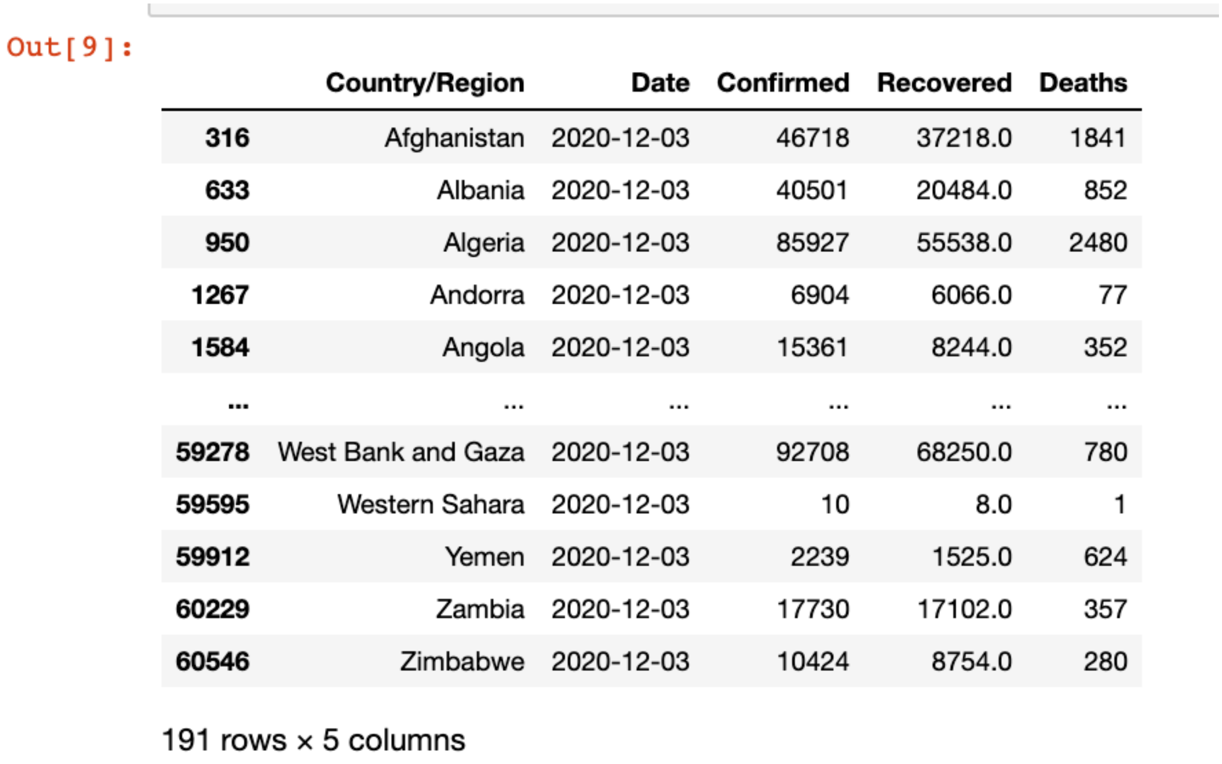
	Country/Region	Date	Confirmed	Recovered	Deaths
0	Afghanistan	2020-01-22	0	0.0	0
1	Afghanistan	2020-01-23	0	0.0	0
2	Afghanistan	2020-01-24	0	0.0	0
3	Afghanistan	2020-01-25	0	0.0	0
4	Afghanistan	2020-01-26	0	0.0	0
...	...	...	...	...	...
60542	Zimbabwe	2020-11-29	9822	8472.0	275
60543	Zimbabwe	2020-11-30	9950	8482.0	276
60544	Zimbabwe	2020-12-01	10129	8643.0	277
60545	Zimbabwe	2020-12-02	10129	8643.0	277
60546	Zimbabwe	2020-12-03	10424	8754.0	280

60547 rows × 5 columns

Figure 2

Compared with *Figure 1*, we can see that in *Figure 2*, there are only 5 columns now, and column Province/State has been removed.

To get the total number of confirmed cases and number of deaths in each country in the last reported day, firstly we call the max function to find the last day, and make column Date in the data frame equal that day. Finally, we get the following result:



Out[9]:

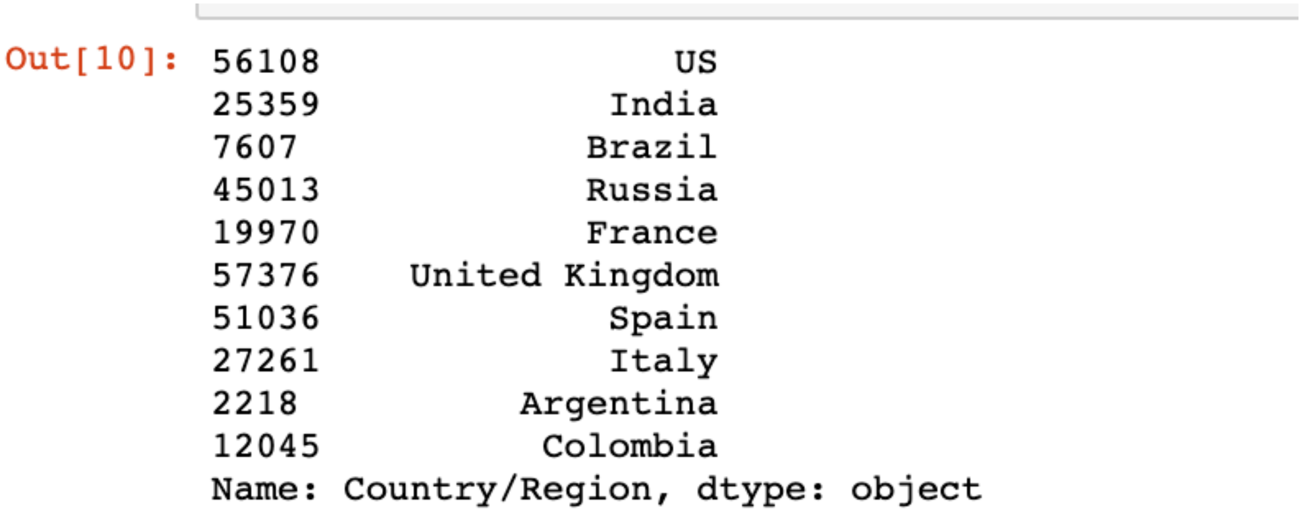
	Country/Region	Date	Confirmed	Recovered	Deaths
316	Afghanistan	2020-12-03	46718	37218.0	1841
633	Albania	2020-12-03	40501	20484.0	852
950	Algeria	2020-12-03	85927	55538.0	2480
1267	Andorra	2020-12-03	6904	6066.0	77
1584	Angola	2020-12-03	15361	8244.0	352
...	...	...	...	...	...
59278	West Bank and Gaza	2020-12-03	92708	68250.0	780
59595	Western Sahara	2020-12-03	10	8.0	1
59912	Yemen	2020-12-03	2239	1525.0	624
60229	Zambia	2020-12-03	17730	17102.0	357
60546	Zimbabwe	2020-12-03	10424	8754.0	280

191 rows × 5 columns

Figure 3

As we can see from Figure 3, the last day of report is 2020-12-03, and we can clearly get that until that day, which country has the largest number of confirmed cases and deaths, and how many people has been cured in each country.

We also want to get the 10 countries with the highest number of confirmed COVID-19 cases. By sorting column Confirmed, we get the following 10 countries with the highest number of confirmed COVID-19 cases:



```
Out[10]: 56108                   US
         25359                India
         7607                Brazil
         45013               Russia
         19970               France
         57376       United Kingdom
         51036                Spain
         27261                Italy
         2218             Argentina
         12045             Colombia
         Name: Country/Region, dtype: object
```

Figure 4

Then we implement matplotlib to plot a graph of the number of confirmed cases over time for the first 20 countries. We will show only 5 graphs here:

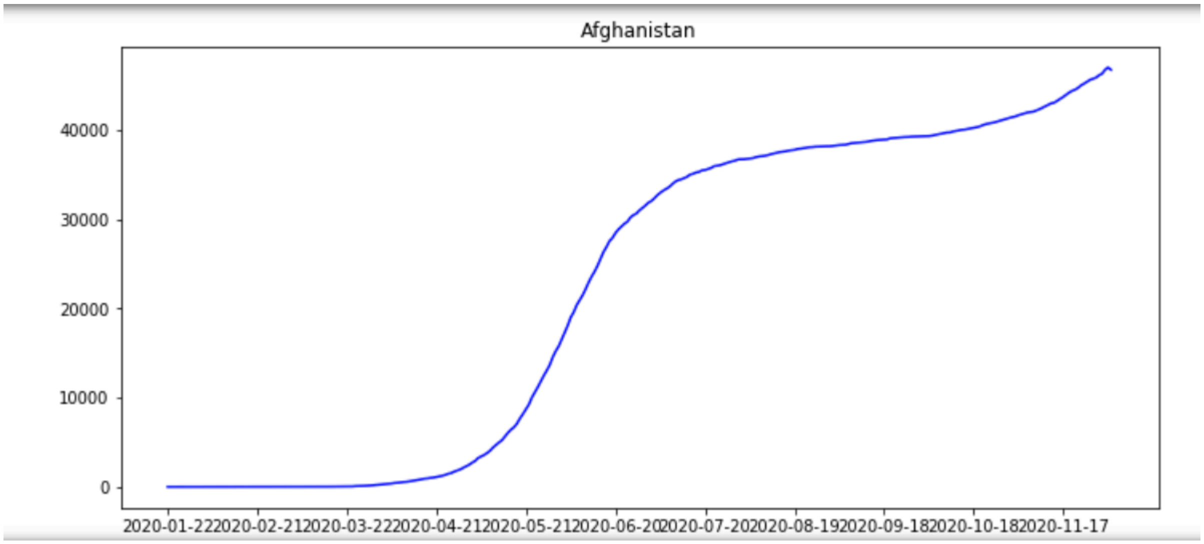


Figure 5

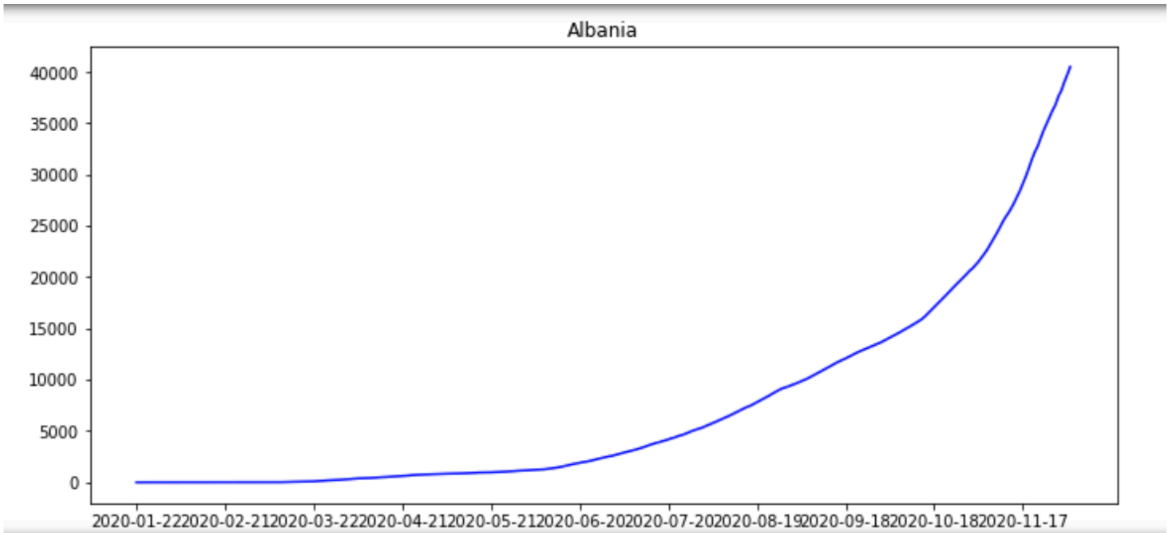


Figure 6

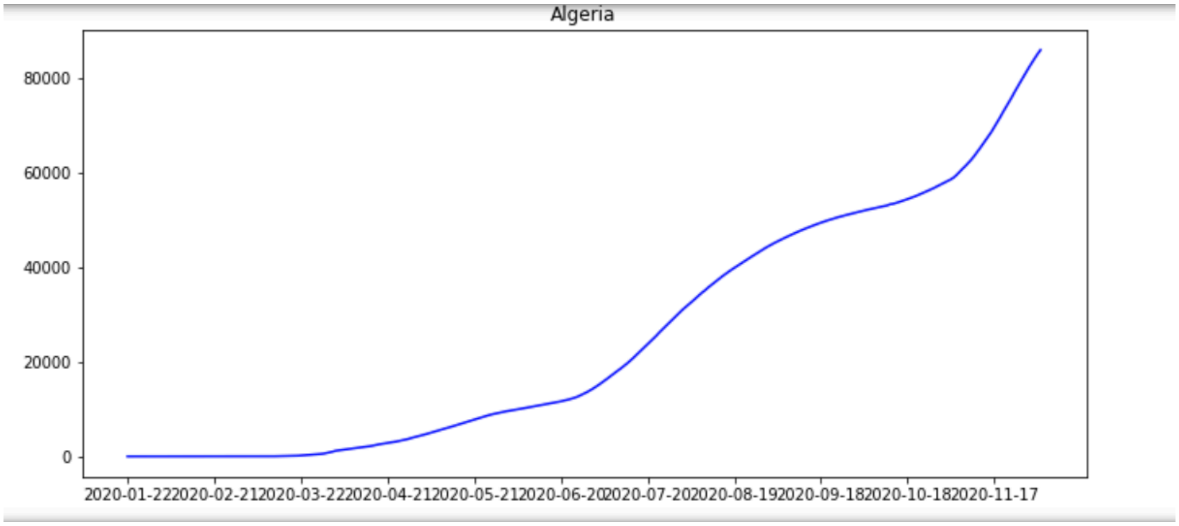


Figure 7

From the 20 graphs we get in Jupyter Notebook, we can see that the following countries present exponential growth in the number of confirmed cases: Albania, Algeria, Andorra, Angola, Argentina, Armenia, Austria, Azerbaijan, Belgium, Belize. The following countries are already leaving exponential growth: Benin, Barbados, Bahamas, Australia, Afghanistan. We can clearly get that COVID-19 is still expanding rapidly in many countries.

,

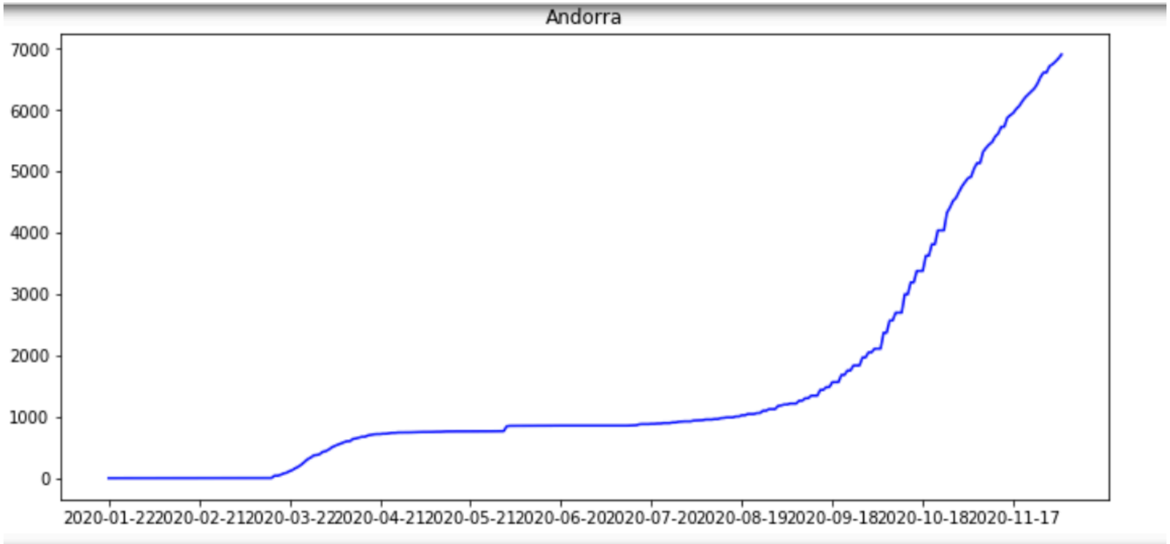


Figure 8

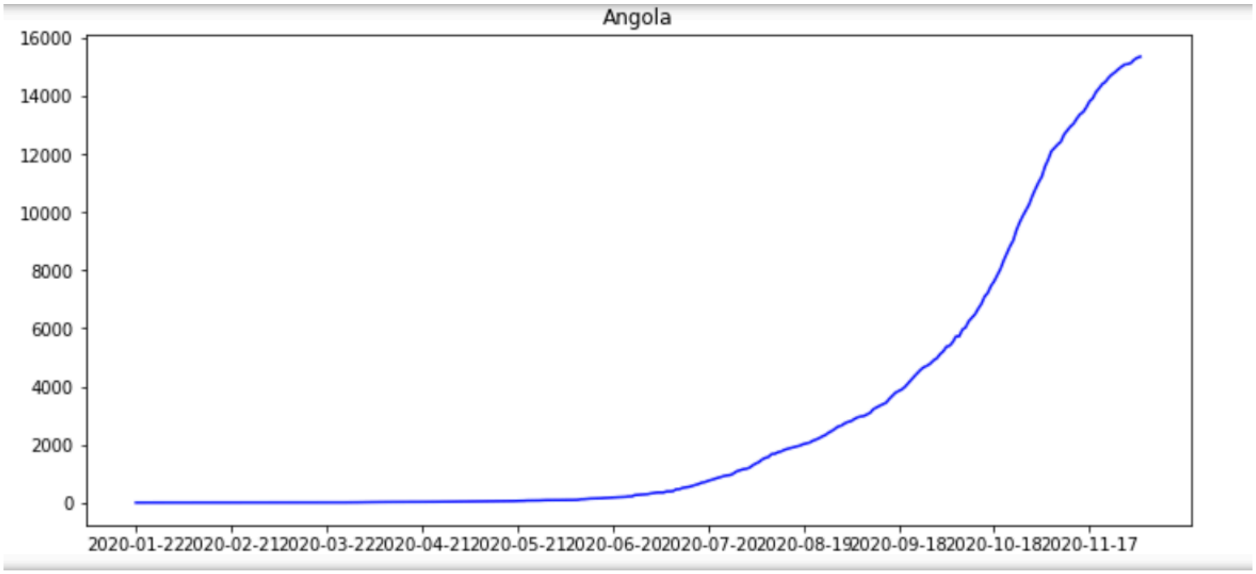
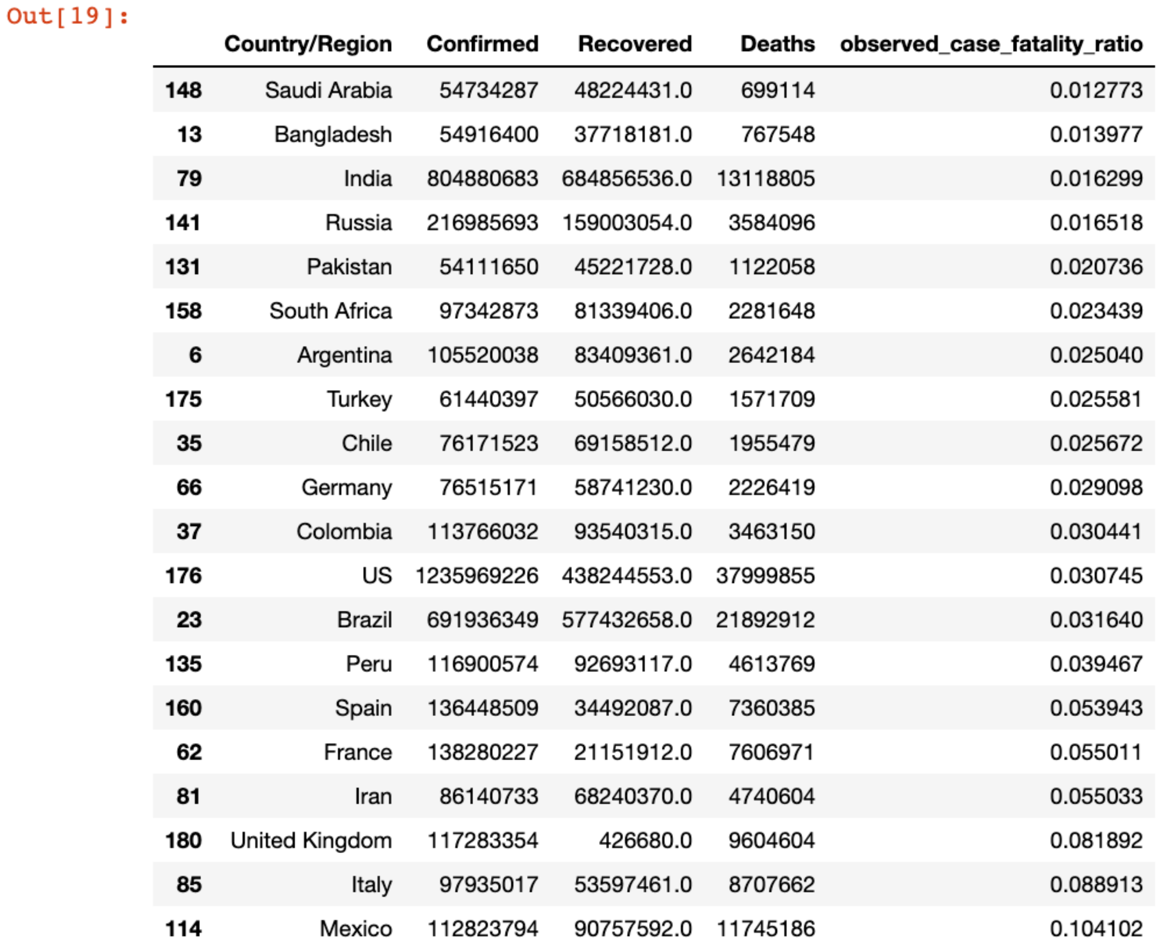


Figure 9

Next, we create a bar plot that shows the number of deaths per 100 confirmed cases (observed case-fatality ratio) for the 20 most affected countries. Firstly, we add a column of observed_case_fatality_ratio to the data frame. Then we plot a bar chart by implementing matplotlib. Here is the new data frame:



Out[19]:

	Country/Region	Confirmed	Recovered	Deaths	observed_case_fatality_ratio
148	Saudi Arabia	54734287	48224431.0	699114	0.012773
13	Bangladesh	54916400	37718181.0	767548	0.013977
79	India	804880683	684856536.0	13118805	0.016299
141	Russia	216985693	159003054.0	3584096	0.016518
131	Pakistan	54111650	45221728.0	1122058	0.020736
158	South Africa	97342873	81339406.0	2281648	0.023439
6	Argentina	105520038	83409361.0	2642184	0.025040
175	Turkey	61440397	50566030.0	1571709	0.025581
35	Chile	76171523	69158512.0	1955479	0.025672
66	Germany	76515171	58741230.0	2226419	0.029098
37	Colombia	113766032	93540315.0	3463150	0.030441
176	US	1235969226	438244553.0	37999855	0.030745
23	Brazil	691936349	577432658.0	21892912	0.031640
135	Peru	116900574	92693117.0	4613769	0.039467
160	Spain	136448509	34492087.0	7360385	0.053943
62	France	138280227	21151912.0	7606971	0.055011
81	Iran	86140733	68240370.0	4740604	0.055033
180	United Kingdom	117283354	426680.0	9604604	0.081892
85	Italy	97935017	53597461.0	8707662	0.088913
114	Mexico	112823794	90757592.0	11745186	0.104102

Figure 10

Here is the bar chart, and the value has been sorted in decrease to gain a clearer picture:

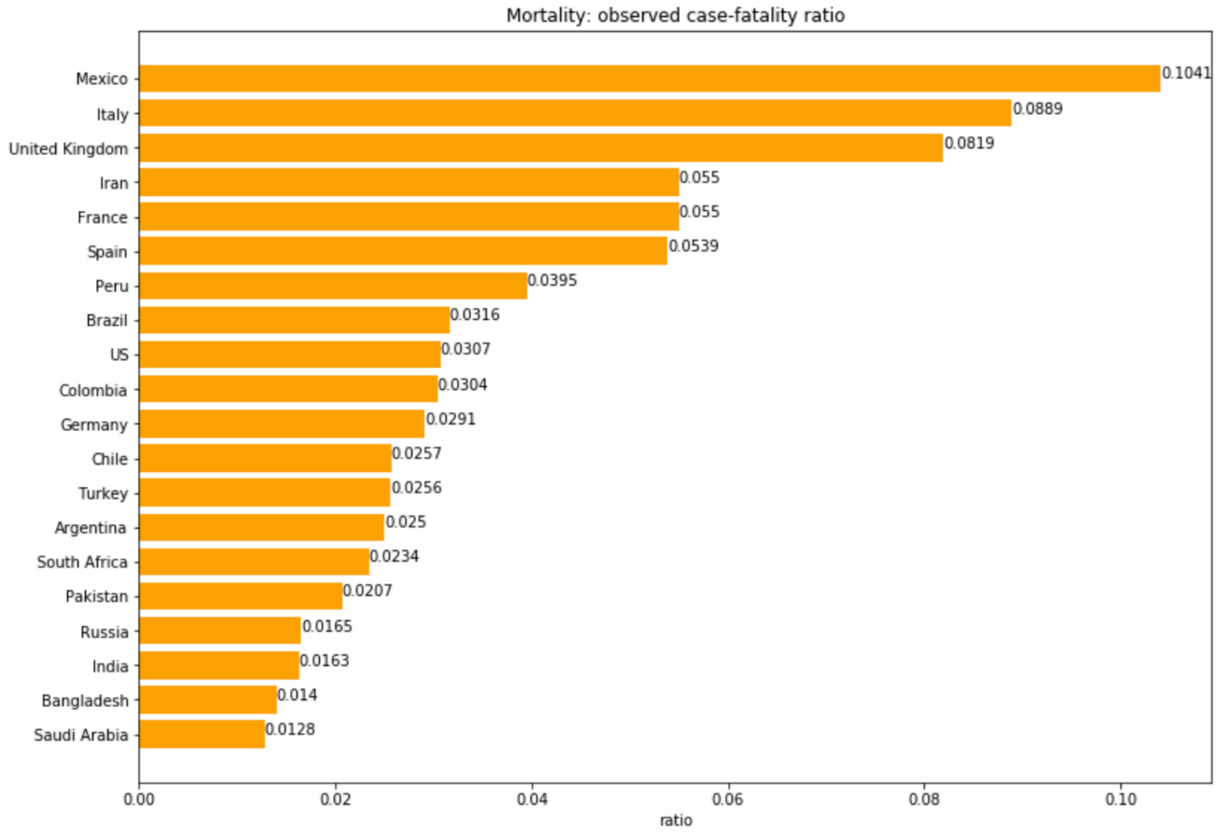
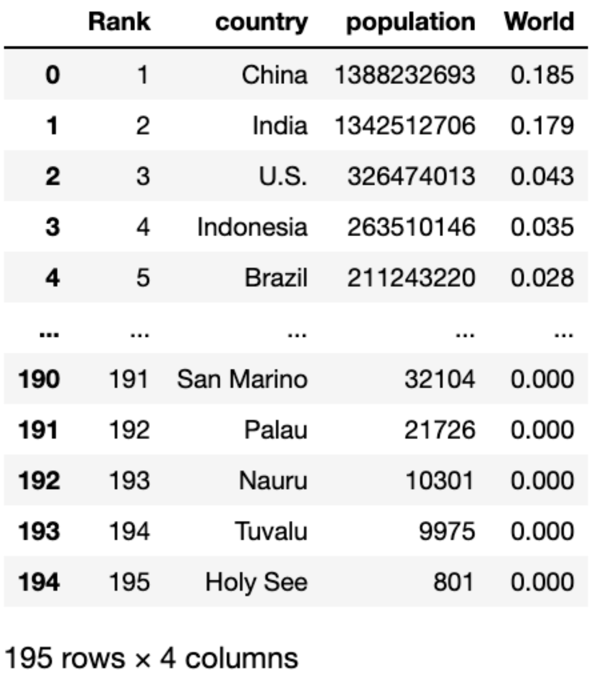


Figure 11

To Compute the ratio between the total number of confirmed cases and the population size for each country, we need to load world population json file and join it with COVID-19 data frame. Here is the json file which has been transformed to a data frame:

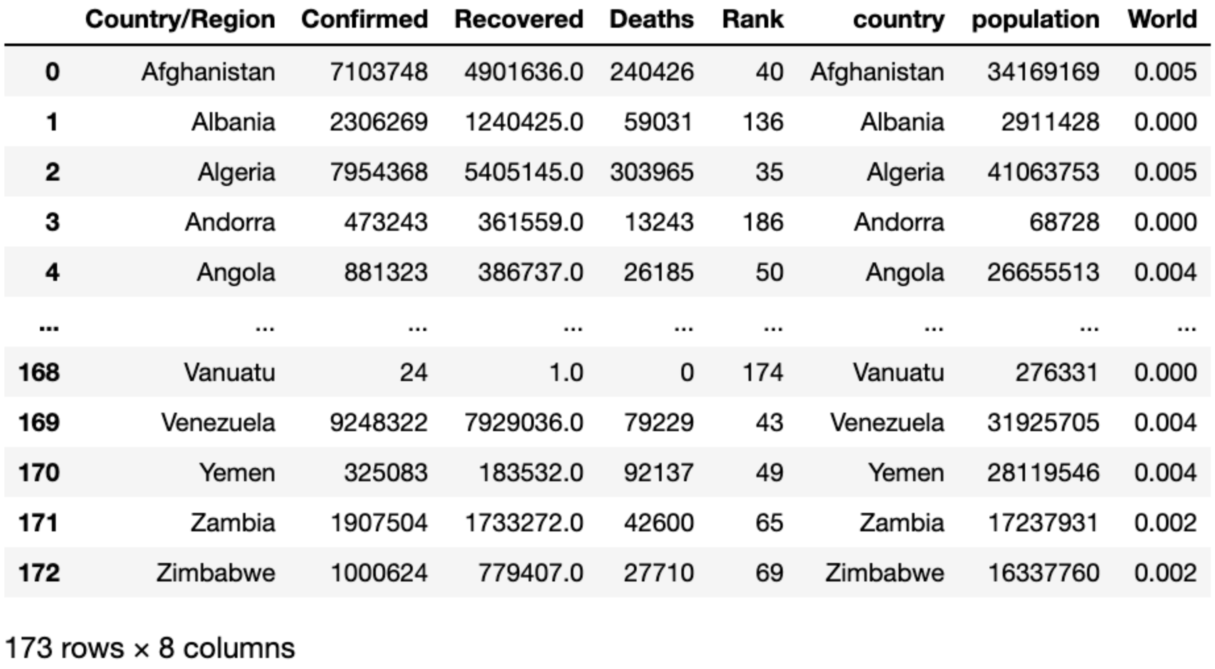


	Rank	country	population	World
0	1	China	1388232693	0.185
1	2	India	1342512706	0.179
2	3	U.S.	326474013	0.043
3	4	Indonesia	263510146	0.035
4	5	Brazil	211243220	0.028
...	...	...	...	...
190	191	San Marino	32104	0.000
191	192	Palau	21726	0.000
192	193	Nauru	10301	0.000
193	194	Tuvalu	9975	0.000
194	195	Holy See	801	0.000

195 rows × 4 columns

Figure 12

Here is the new data frame that has been joined with world population data set:

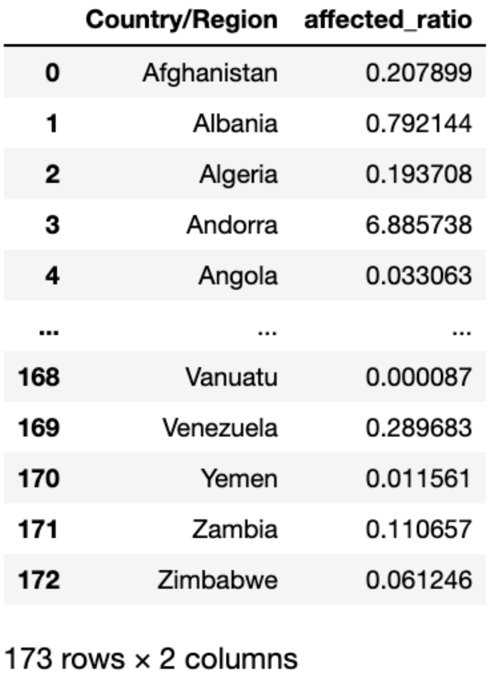


	Country/Region	Confirmed	Recovered	Deaths	Rank	country	population	World
0	Afghanistan	7103748	4901636.0	240426	40	Afghanistan	34169169	0.005
1	Albania	2306269	1240425.0	59031	136	Albania	2911428	0.000
2	Algeria	7954368	5405145.0	303965	35	Algeria	41063753	0.005
3	Andorra	473243	361559.0	13243	186	Andorra	68728	0.000
4	Angola	881323	386737.0	26185	50	Angola	26655513	0.004
...	...	...	...	...	...	...	...	...
168	Vanuatu	24	1.0	0	174	Vanuatu	276331	0.000
169	Venezuela	9248322	7929036.0	79229	43	Venezuela	31925705	0.004
170	Yemen	325083	183532.0	92137	49	Yemen	28119546	0.004
171	Zambia	1907504	1733272.0	42600	65	Zambia	17237931	0.002
172	Zimbabwe	1000624	779407.0	27710	69	Zimbabwe	16337760	0.002

173 rows × 8 columns

Figure 13

Here is the new data frame with ratio between the total number of confirmed cases and the population size for each country:

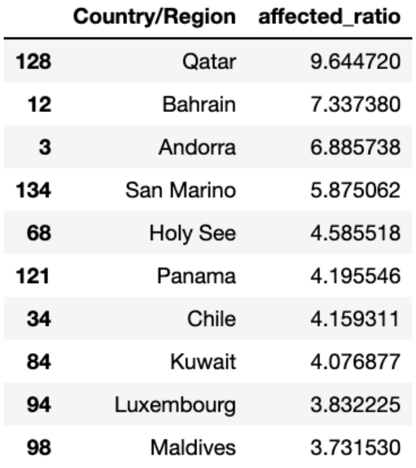


	Country/Region	affected_ratio
0	Afghanistan	0.207899
1	Albania	0.792144
2	Algeria	0.193708
3	Andorra	6.885738
4	Angola	0.033063
...	...	...
168	Vanuatu	0.000087
169	Venezuela	0.289683
170	Yemen	0.011561
171	Zambia	0.110657
172	Zimbabwe	0.061246

173 rows × 2 columns

Figure 14

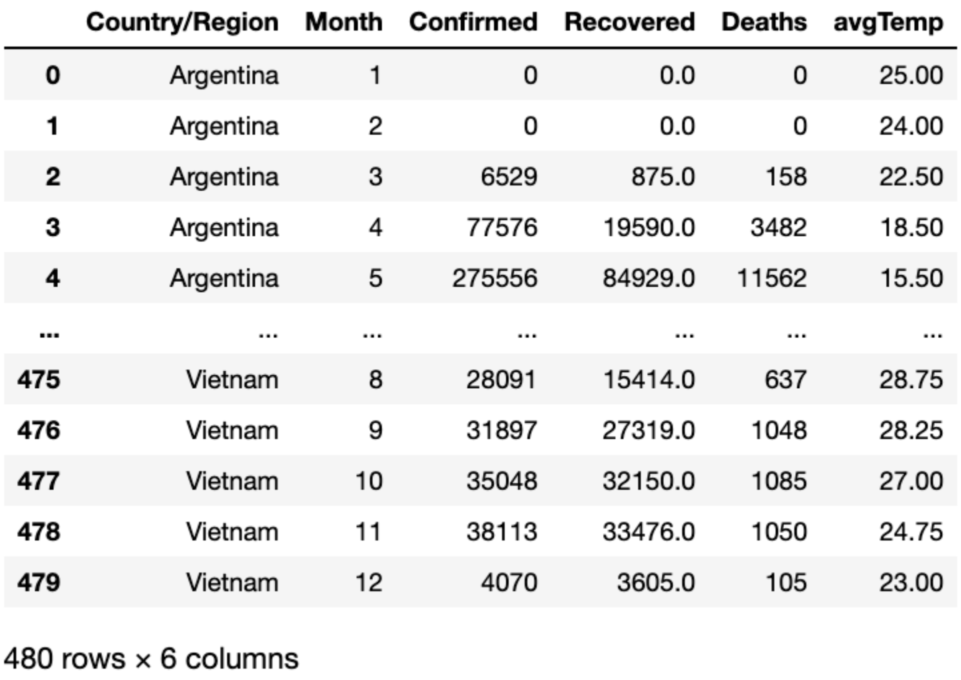
Based on the above results, we can easily get what are the 10 countries with the highest number of confirmed COVID-19 cases per capita:



	Country/Region	affected_ratio
128	Qatar	9.644720
12	Bahrain	7.337380
3	Andorra	6.885738
134	San Marino	5.875062
68	Holy See	4.585518
121	Panama	4.195546
34	Chile	4.159311
84	Kuwait	4.076877
94	Luxembourg	3.832225
98	Maldives	3.731530

Figure 15

Next, we would like to test the hypothesis that the spread of the virus is slowed down by warm weather. To do this, we should firstly normalize data and clean data. We import climate json file by importing json_normalize from pandas.io.json, execute two for loops, and group by country and month to get the average monthly temperature for each country. Finally we merge monthly number of confirmed cases with monthly average temperature to get a new DataFrame and do analysis. Here is the final DataFrame we get:



	Country/Region	Month	Confirmed	Recovered	Deaths	avgTemp
0	Argentina	1	0	0.0	0	25.00
1	Argentina	2	0	0.0	0	24.00
2	Argentina	3	6529	875.0	158	22.50
3	Argentina	4	77576	19590.0	3482	18.50
4	Argentina	5	275556	84929.0	11562	15.50
...	...	...	...	...	...	...
475	Vietnam	8	28091	15414.0	637	28.75
476	Vietnam	9	31897	27319.0	1048	28.25
477	Vietnam	10	35048	32150.0	1085	27.00
478	Vietnam	11	38113	33476.0	1050	24.75
479	Vietnam	12	4070	3605.0	105	23.00

480 rows × 6 columns

Figure 16

Then select a few countries to plot graph between the two factors and analyze their correlation.

Here we plot graphs for three countries and get the following results:

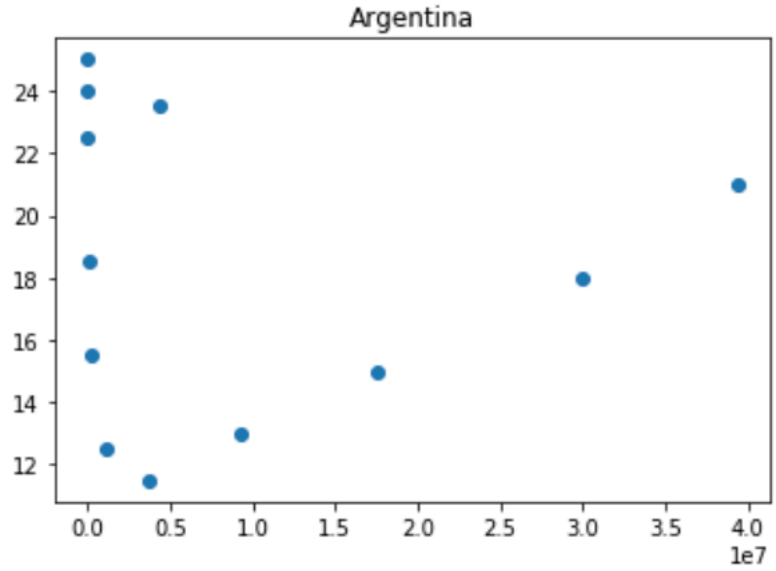


Figure 17

The correlation is about -0.03 which indicates that the monthly confirmed cases has negative correlation with monthly average temperature.

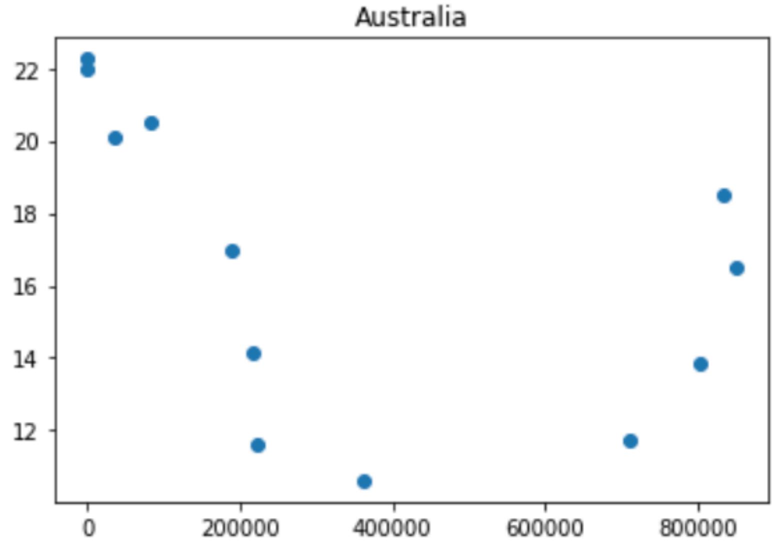


Figure 18

The correlation is about -0.46 which indicates that the monthly confirmed cases has negative correlation with monthly average temperature.

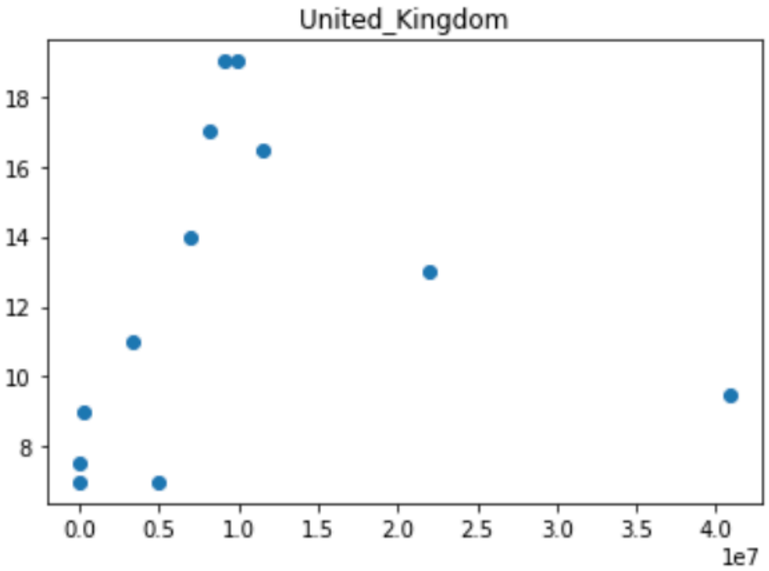
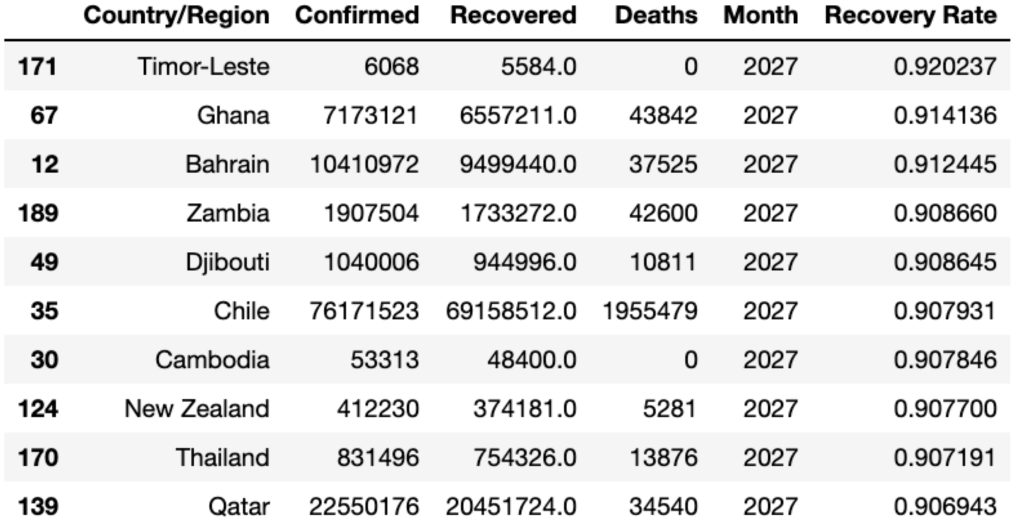


Figure 19

The correlation is about 0.13 which indicates that the monthly confirmed cases has negligible correlation with monthly average temperature.

From what we analyzed, we can conclude the hypothesis that the spread of the virus is slowed down by warm weather is not true.

Finally, we'd like to get the 10 countries with the highest recovery rate. Here are the 10 countries we get:



	Country/Region	Confirmed	Recovered	Deaths	Month	Recovery Rate
171	Timor-Leste	6068	5584.0	0	2027	0.920237
67	Ghana	7173121	6557211.0	43842	2027	0.914136
12	Bahrain	10410972	9499440.0	37525	2027	0.912445
189	Zambia	1907504	1733272.0	42600	2027	0.908660
49	Djibouti	1040006	944996.0	10811	2027	0.908645
35	Chile	76171523	69158512.0	1955479	2027	0.907931
30	Cambodia	53313	48400.0	0	2027	0.907846
124	New Zealand	412230	374181.0	5281	2027	0.907700
170	Thailand	831496	754326.0	13876	2027	0.907191
139	Qatar	22550176	20451724.0	34540	2027	0.906943

Figure 20

V. Conclusions

After doing the above analysis and visualizing the results, we can learn that COVID-19 is a disease which spreads very rapidly, and we can see from our visualization that in most countries there is an exponential pattern of outbreaks. By doing a hypothesis test, although there is no absolute correlation between the rise in cases and temperature, a second wave is under way as air temperatures fall and humidity increases. As for the mortality rate of COVID-19, it varies greatly from the country to country, which has a great relationship with the level of medical development and economic level of each country, but the mortality rate is not high overall compared with that of other epidemics. After analyzing the recovery rate in each country, we can see that most countries have over 90 percent of recovery, which is the positive side.

In this study, we have not studied and analyzed the relationship between COVID-19 and geographical location, which is the limits of this study. The good thing about this report is that it gives people a basic understanding of COVID-19 in terms of transmission rate, mortality rate, recovery rate etc., and gives people a picture of COVID-19 transmission in every country in the world.