MSDS 422 Assignment 1

Husein Adenwala

Data preparation

I am using the preprocessed dataset provided by the course github page. The original data on ECDC's website had some erroneous negative cases and deaths and NA values, which had already been removed from the dataset I used. I did not have to strip out any other datapoints. Additionally, the date column was formatted to the correct format in the dataset, (year month and day) per the convention.

In addition to these items, I added two columns to prepare for analysis: 1) cumulative death by country and 2) cumulative cases by country. This will be useful in showing the trend in the data.

This dataset had 38492 rows and 10 columns. The data are categorized into 210 countries and 5 continents (North and South America are grouped as one). Since there are only 195 countries, 15 of these may be regions or territories that are keeping their own statistics.

(df1)

Date	Day	Month	Year	Cases	Deaths	Country	Population	Continent	CumulativeNumberPer100KCases	cumsum_cases	cumsum_death
2020-08-25	25	8	2020	71	10	Afghanistan	38041757	Asia	2.670749	38070	1397
2020-08-24	24	8	2020	0	0	Afghanistan	38041757	Asia	2.484112	37999	1387
2020-08-23	23	8	2020	105	2	Afghanistan	38041757	Asia	2.484112	37999	1387
2020-08-22	22	8	2020	38	0	Afghanistan	38041757	Asia	2.310619	37894	1385
2020-08-21	21	8	2020	97	2	Afghanistan	38041757	Asia	2.415766	37856	1385

Once I added cumulative cases and cumulative death by country columns, I created a separate data frame with aggregate death and cases per country. I added a column for death rate as a percentage of reported cases and death and cases per million (population) to standardize the data. This new data frame has 210 rows (# of countries/regions) and 6 columns (see below).

(df2)

Country	Population	Deaths	Cases	Death_per_million	Death_pct_cases	Cases_per_million
Yemen	29161922	555	1916	19.031667	28.966597	65.702117
Italy	60359546	35503	260594	588.191966	13.623875	4317.361830
United_Kingdom	66647112	41433	326614	621.677350	12.685617	4900.647458
France	67012883	30528	246386	455.554195	12.390314	3676.696017
Belgium	11455519	9996	81998	872.592503	12.190541	7157.947187
Hungary	9772756	613	5191	62.725397	11.808900	531.170532
Mexico	127575529	60800	563705	476.580426	10.785783	4418.598178
Netherlands	17282163	6193	67062	358.346348	9.234738	3880.417052
Jersey	107796	32	364	296.857026	8.791209	3376.748673
Spain	46937060	32712	407606	696.933297	8.025397	8684.097385

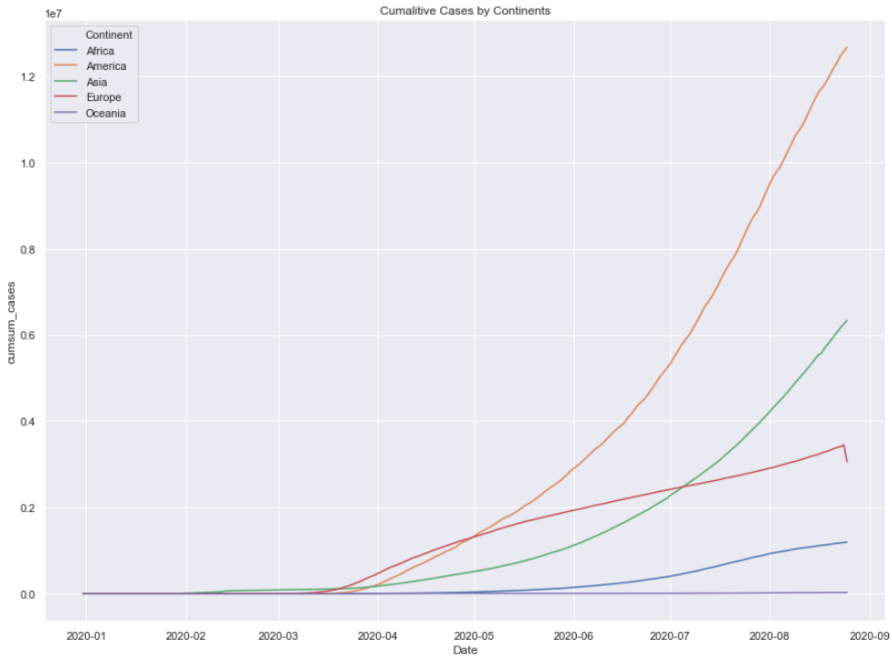
Data exploration

Looking at the cumulative death and cumulative cases by date and country, we can see that max cumulative death is 177,279 (USA) and 5.7 million max total number of cases (USA) by a country. The below table provides the summary of the data frame, including the min, max, standard deviation and data types.

	Day	Month	Year	Cases	Deaths	Population	CumulativeNumberPer100KCases	cumsum_cases	cumsum_death
count	38492.000000	38492.000000	38492.000000	38492.000000	38492.000000	3.849200e+04	38492.000000	3.849200e+04	38492.000000
mean	15.810558	5.192585	2019.998259	615.382937	21.247376	4.384876e+07	27.128374	3.617301e+04	1662.431518
std	8.694756	1.995496	0.041685	3815.217442	125.237191	1.598573e+08	67.447288	2.281671e+05	9212.465554
min	1.000000	1.000000	2019.000000	0.000000	0.000000	8.150000e+02	-1.262589	0.000000e+00	0.000000
25%	8.000000	4.000000	2020.000000	0.000000	0.000000	1.394969e+06	0.060058	3.000000e+01	0.000000
50%	16.000000	5.000000	2020.000000	7.000000	0.000000	8.519373e+06	2.882863	7.030000e+02	12.000000
75%	23.000000	7.000000	2020.000000	126.000000	2.000000	3.036604e+07	19.221559	7.377000e+03	149.000000
max	31.000000	12.000000	2020.000000	78427.000000	4928.000000	1.433784e+09	1058.225943	5.740909e+06	177279.000000

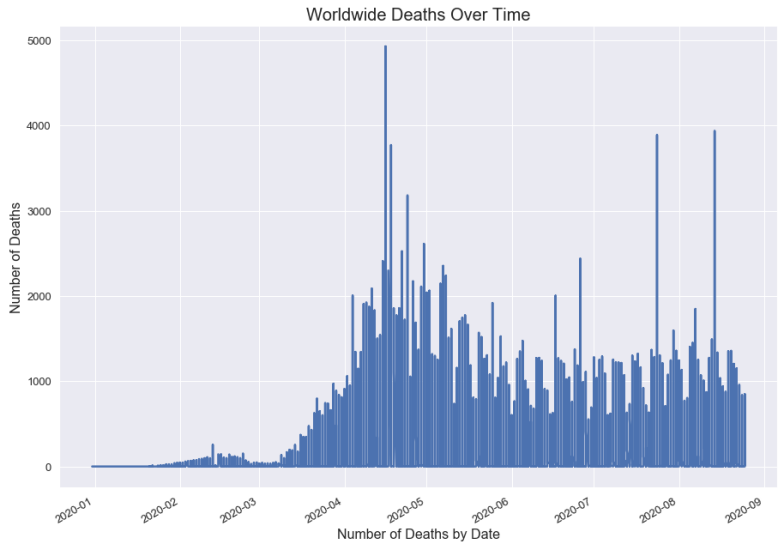
```
Date                          datetime64[ns]
Day                                    int64
Month                                  int64
Year                                   int64
Cases                                  int64
Deaths                                 int64
Country                               object
Population                             int64
Continent                             object
CumulativeNumberPer100KCases         float64
cumsum_cases                           int64
cumsum_death                           int64
```

Cumulative cases grouped by continent shows the trend by different continents. Some continents (e.g. Oceania, Africa and Europe) are doing better and some are doing worse (e.g. America, Asia).

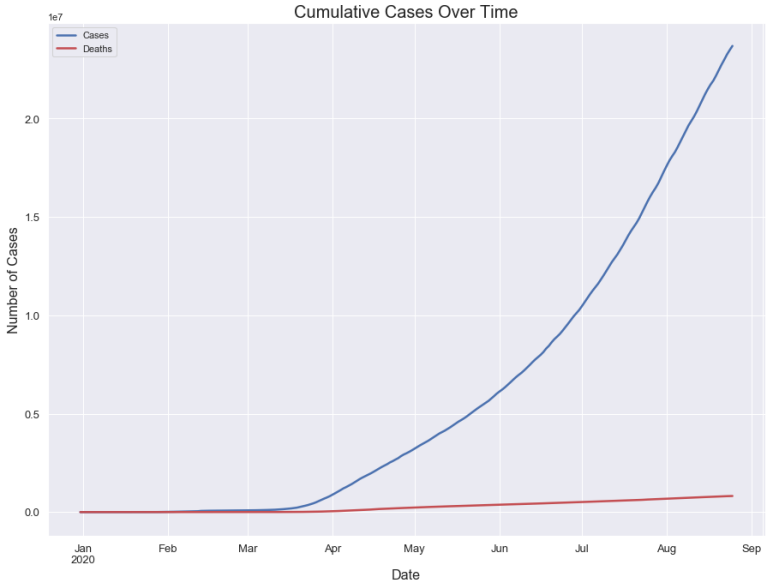



```
Continent
Africa       1196602
America     12660833
Asia         6334181
Europe       3449313
Oceania        27978
```

The world wide number of deaths by day shows some unusual spikes which might be due to infrequent reporting by some countries.



The cumulative cases and death graph shows a higher rate of increase in number of cases compared to number of deaths, which reflects more testing, which can reduce or lower the death rate.



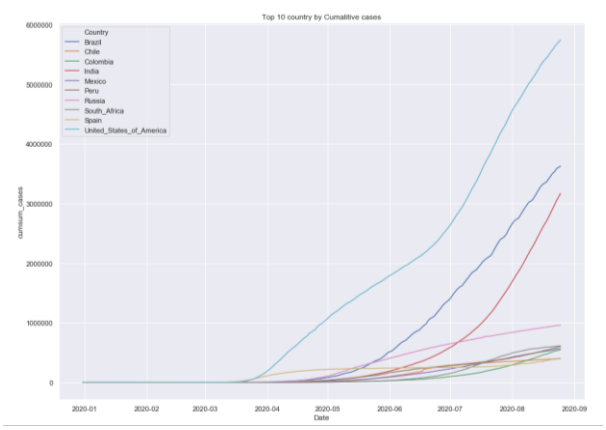
Date	Cases
2020-08-25	23687320

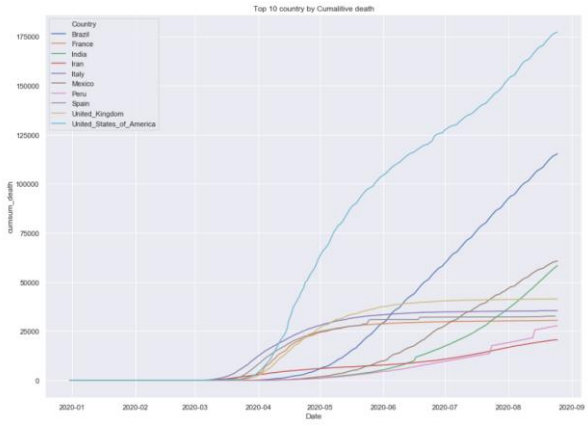
Date	Deaths
2020-08-25	817854

Data visualization

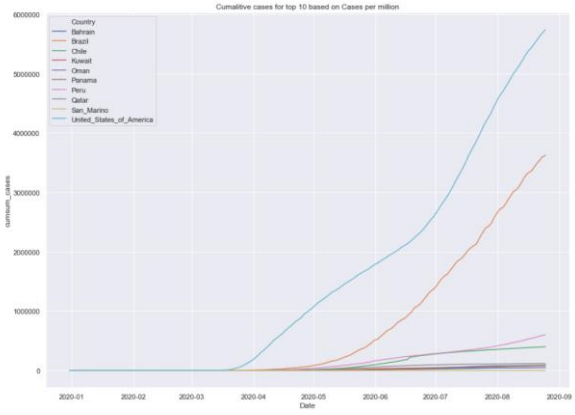
For the data visualization part, I wanted to see the trend by country of COVID19 related cases and death. To do this, I selected the top 10 countries by total number of death and top 10 countries by total number of cases.

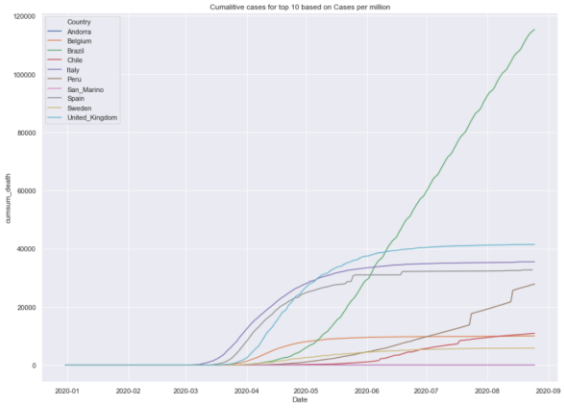
Looking at cumulative cases and deaths for the top 10 countries for highest COVID19 cases and deaths, Russia is one of the top 10 for cases, but not for deaths and Iran is one of the top 10 in deaths but not in cases.





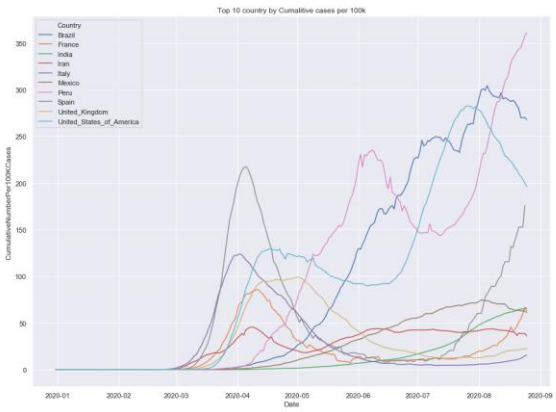
I want to add a caveat here that total number of cases and death might not indicate the most affected countries as it does not account for differences in population. For that, we should look at death and cases per million. I have used cases per million and death per million in the second data frame (df2) to isolate the top 10 countries (in cases and death per million), then compare them to each other.

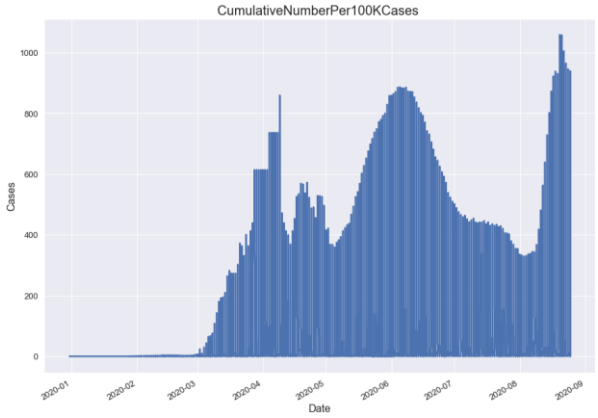




We can see the list of countries change from the previous two graphs and most notably, the US is not in the top 10 deaths per millon list.

The below graphs shows the 14-day cumulative cases for the top 10 countries by total number of cases and 14-day cumulative cases for all the countries together:

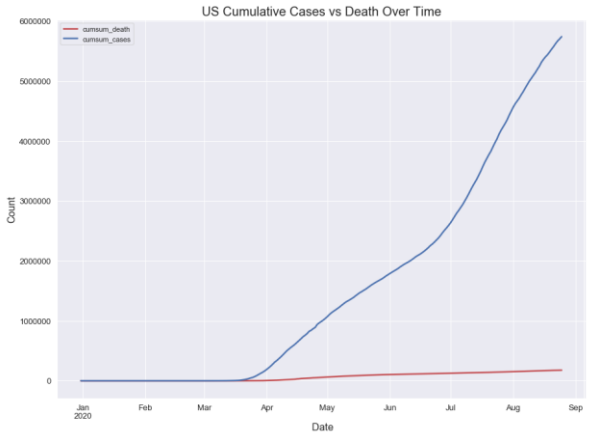


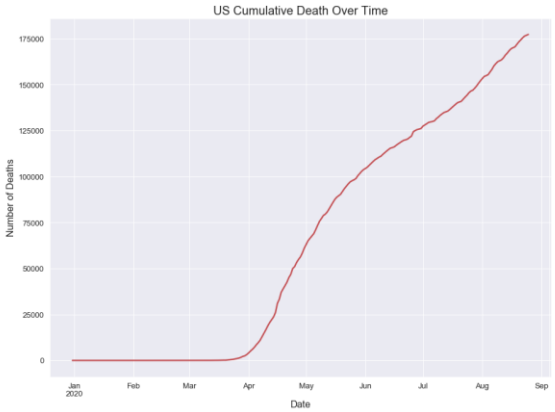


Looking at the 14-day cumulative cases per 100,000 people, we can see that there is a rise in global COVID19 cases, and this makes sense given the current new reporting on COVID19 cases worldwide. Only some countries in the top 10 total number of cases show a downward trend.

Data scaling and comparisons

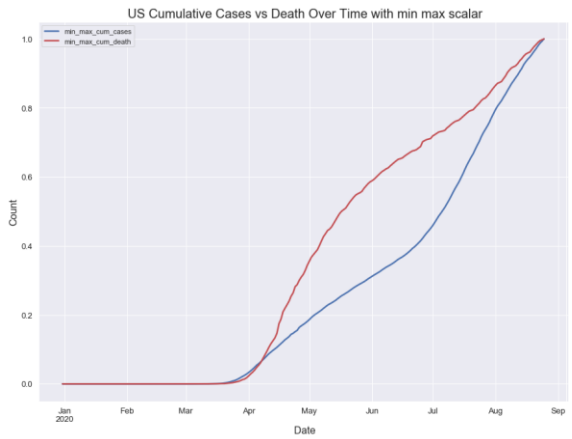
In addition to visualizing and comparing the cumulative cases around the world, I chose to compare cumulative COVID19 cases and death in the US. To do this, I used min, max and standard scaling to normalize the data for comparison. The two charts below show the cases and death as per the original column, prior to scaling.

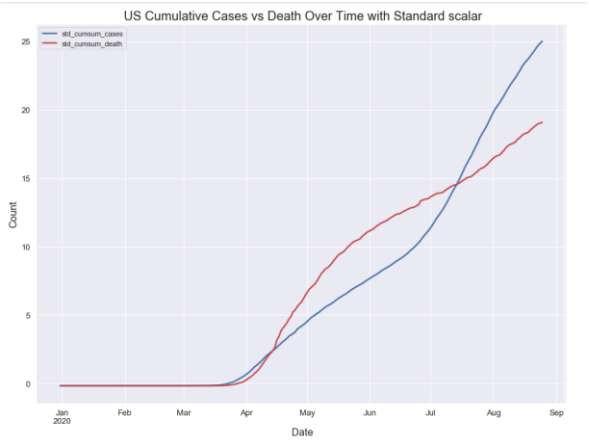




After doing the min, max and standard scaling, we can see that the distribution has the same shape but the scale has changed and that makes it comparable. We can see that the number of cases in the US is increasing at a high rate whereas the curve of the death rate shows a decrease in the rate of cumulative death.

This trend indicates that there is an increase in the number of cases and increase in testing capacity as the death rate is slowing down.





The min, max and standard scaling don't perform well with outliers and it is useful when features have a normal distribution. This is not the best data fit for these functions. Boxplot for cases, deaths or cumulative cases and deaths show that the data is extremely right skewed and has many outliers.

Insights from analysis

Looking at the percentage of cases that result in death, we can assume that some countries are under-testing or -reporting COVID cases and some may be under-reporting deaths caused by COVID as well. In the data visualization above, we find Iran is in the top 10 in number of COVID-related deaths in the world but not in the top 10 COVID cases, whereas Russia is in the top 10 in reporting cases but not for COVID-related death. Total Death as a percentage of total cases ("Death_pct_cases") would make a good indicator for over/underreporting or low/high testing.

Country	Population	Deaths	Cases	Death_per_million	Death_pct_cases
Iran	82913893	20643	358905	248.969132	5.751661
Russia	145872260	16448	961493	112.756188	1.710673

In the US, we can see a positive trend of increase in rate of testing and decrease in the rate of COVID-related death. The US still has the highest number of COVID-related deaths and cases; however, looking at the 14-day moving average cases per 100k, the US is on a decline.

The 14-day moving average also shows some interesting insights into countries such as Peru and Mexico, who are on an upward trend of COVID cases; it seems that worldwide, cases of COVID19 went up in August.

Looking at the cumulative cases by continent, we can see that countries in Europe, Oceania and Africa have managed to reduce the rate of spread of COVID. There is a slight caveat with the African countries, because a lot of them have 0 cases, which makes it seem as though there is no data coming from those countries, not necessarily that there is no virus. Countries in Oceania, on the other hand, seem to have

reduced the spread of COVID significantly and their COVID-related policies may be useful models for the rest of the world.

The US has the highest reported cases and deaths but looking at deaths per million and cases per million, which normalizes the data for total population, we can see that the US ranks 11th and 8th respectively.

	Country	Population	Deaths	Cases	Death_per_million	Death_pct_cases	Cases_per_million
0	San_Marino	34453	42	744	1219.052042	5.645161	21594.636171
1	Belgium	11455519	9996	81998	872.592503	12.190541	7157.947187
2	Peru	32510462	27813	600438	855.509220	4.632119	18469.070049
3	Spain	46937060	32712	407606	696.933297	8.025397	8684.097385
4	Andorra	76177	53	1060	695.748060	5.000000	13914.961209
5	United_Kingdom	66647112	41433	326614	621.677350	12.685617	4900.647458
6	Italy	60359546	35503	260594	588.191966	13.623875	4317.361830
7	Chile	18952035	10916	399568	575.980363	2.731951	21083.118515
8	Sweden	10230185	5813	86721	568.220418	6.703105	8476.972802
9	Brazil	211049519	115309	3622861	546.359928	3.182816	17165.928722
10	United_States_of_America	329064917	177279	5740909	538.735644	3.087995	17446.129026

	Country	Population	Deaths	Cases	Death_per_million	Death_pct_cases	Cases_per_million
0	Qatar	2832071	194	117266	68.501107	0.165436	41406.447790
1	Bahrain	1641164	184	49330	112.115547	0.372998	30057.934490
2	San_Marino	34453	42	744	1219.052042	5.645161	21594.636171
3	Chile	18952035	10916	399568	575.980363	2.731951	21083.118515
4	Panama	4246440	1906	87485	448.846563	2.178659	20601.963056
5	Kuwait	4207077	518	80960	123.125866	0.639822	19243.764733
6	Peru	32510462	27813	600438	855.509220	4.632119	18469.070049
7	United_States_of_America	329064917	177279	5740909	538.735644	3.087995	17446.129026
8	Brazil	211049519	115309	3622861	546.359928	3.182816	17165.928722
9	Oman	4974992	637	84509	128.040407	0.753766	16986.760984

Appendix:

Code for standard Scalar

```
sc=StandardScaler()
```

```
covid19_dfA["std_cumsum_cases"],covid19_dfA["std_cumsum_death"]=\
sc.fit_transform(covid19_dfA[["cumsum_cases","cumsum_death"]])[:,0],sc.fit_transform(covid19_dfA[["cumsum_cases","cumsum_death"]
```

```
fig, ax = plt.subplots()
sns.set(rc={'figure.figsize':(15, 11)})
US_daily['std_cumsum_cases'].plot(ax= ax, linewidth = 2.5)
US_daily['std_cumsum_death'].plot(ax= ax, linewidth = 2.5, color = 'r')

plt.title(' US Cumulative Cases vs Death Over Time with Standard scalar', fontsize = 20)
plt.xlabel(' Date', fontsize = 16)
plt.xticks(fontsize = 13)
plt.ylabel('Count', fontsize = 16)
plt.yticks(fontsize = 13)
leg = ax.legend()

plt.show()
```

Code for min max scaler and graph

```
covid19_dfA["min_max_cum_cases"] = minmax_scale(covid19_dfA["cumsum_cases"])
covid19_dfA["min_max_cum_death"] = minmax_scale(covid19_dfA["cumsum_death"])
```

```
# cumsum death vs cases wtih min max scalar
fig, ax = plt.subplots()
sns.set(rc={'figure.figsize':(15, 11)})
US_daily['min_max_cum_cases'].plot(ax= ax, linewidth = 2.5)
US_daily['min_max_cum_death'].plot(ax= ax, linewidth = 2.5, color = 'r')

plt.title(' US Cumulative Cases vs Death Over Time with min max scalar', fontsize = 20)
plt.xlabel(' Date', fontsize = 16)
plt.xticks(fontsize = 13)
plt.ylabel('Count', fontsize = 16)
plt.yticks(fontsize = 13)
leg = ax.legend()

plt.show()
```

Low death rate per cases

Country	Population	Deaths	Cases	Death_per_million	Death_pct_cases	mi
Laos	7169456	0	22	0.000000	0.000000	
Grenada	112002	0	24	0.000000	0.000000	
Cambodia	16486542	0	273	0.000000	0.000000	
Greenland	56660	0	14	0.000000	0.000000	
New_Caledonia	282757	0	23	0.000000	0.000000	
Timor_Leste	1293120	0	26	0.000000	0.000000	
Bonaire, Saint Eustatius and Saba	25983	0	13	0.000000	0.000000	
Dominica	71808	0	20	0.000000	0.000000	
Bhutan	763094	0	156	0.000000	0.000000	
Gibraltar	33706	0	248	0.000000	0.000000	
Holy_See	815	0	12	0.000000	0.000000	
Seychelles	97741	0	132	0.000000	0.000000	
Eritrea	3497117	0	306	0.000000	0.000000	
French_Polynesia	279285	0	310	0.000000	0.000000	
Saint_Vincent_and_the_Grenadines	110593	0	58	0.000000	0.000000	
Saint_Kitts_and_Nevis	52834	0	17	0.000000	0.000000	
Falkland_Islands_(Malvinas)	3372	0	13	0.000000	0.000000	
Saint_Lucia	182795	0	26	0.000000	0.000000	
Faroe_Islands	48677	0	410	0.000000	0.000000	
Anguilla	14872	0	3	0.000000	0.000000	
Mongolia	3225166	0	298	0.000000	0.000000	
Singapore	5804343	27	56404	4.651689	0.047869	
Western_Sahara	582458	1	766	1.716862	0.130548	
Qatar	2832071	194	117266	68.501107	0.165436	
Botswana	2303703	3	1562	1.302251	0.192061	
Burundi	11530577	1	430	0.086726	0.232558	
Bahrain	1641164	184	49330	112.115547	0.372998	
Maldives	530957	27	6912	50.851576	0.390625	
Sri_Lanka	21323734	12	2959	0.562753	0.405542	
Rwanda	12626938	14	3306	1.108741	0.423472	
Aruba	106310	7	1628	65.845170	0.429975	
Nepal	28608715	157	32678	5.487838	0.480446	
Iceland	356991	10	2073	28.011911	0.482393	
Cayman_Islands	64948	1	205	15.396933	0.487805	
Turks_and_Caicos_islands	38194	2	383	52.364246	0.522193	
United_Arab_Emirates	9770526	376	67282	38.483087	0.558842	
Palestine	4981422	147	25588	29.509646	0.574488	
Guinea	12771246	54	9013	4.228248	0.599135	
Malta	493559	10	1667	20.261002	0.599880	
Ghana	30417858	263	43622	8.646237	0.602907	

One of codes for top 10 countries

```
top10 = covid19_dfA.loc[covid19_dfA["Country"].isin(covid19_deaths.nlargest(10, ["Death_per_million"])["Country"])]
```

```
#covid19_deaths.nlargest(10,["Cases_per_million"])
```

```
sns.lineplot(x= "Date", y = "cumsum_cases", data = top10, hue = "Country").set_title("Top 10 country by Cumalitive cases by Case
```