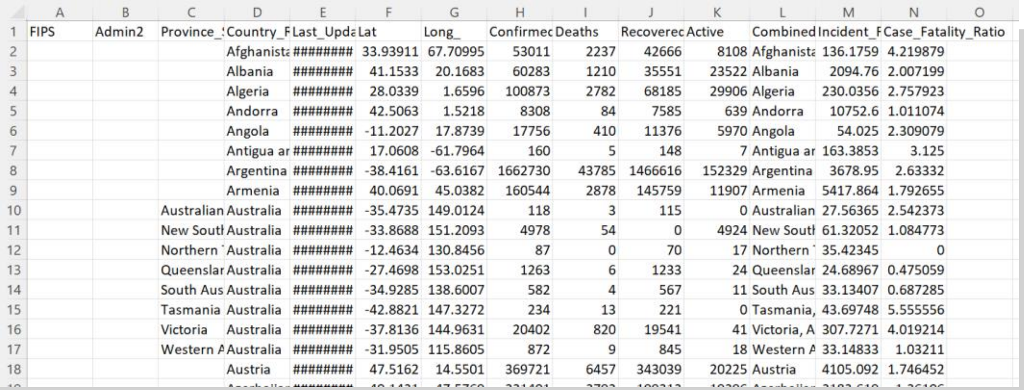
Spark Project

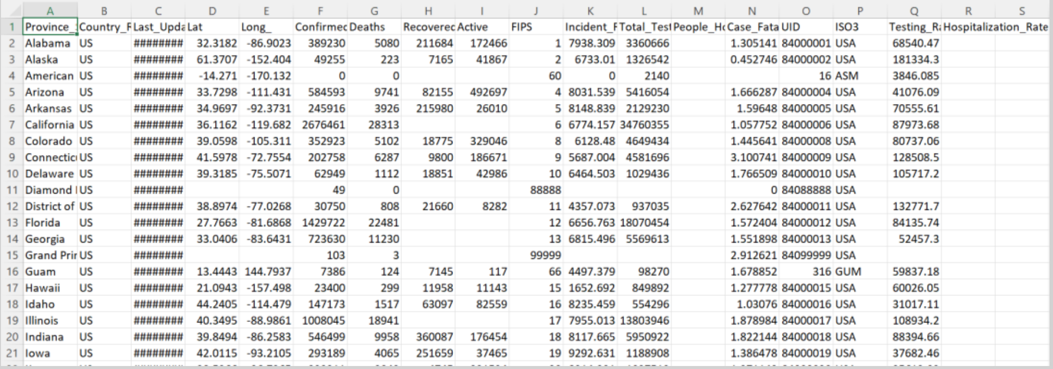
4/5/2022

Dataset - Johns Hopkins CSSE Data

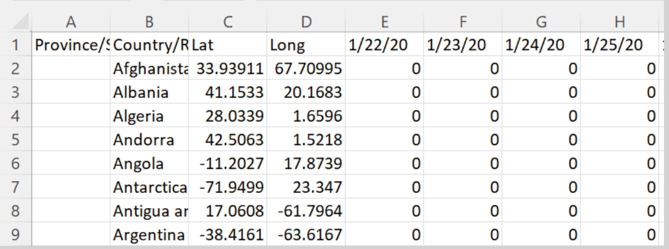
- Download dataset from:
  - https://github.com/CSSEGISandData/COVID-19
- Data Folder Structure:
  - archived_data/
    - archived_daily_case_updates/
      - csv for 01-21-2020 to 02-14-2020
    - archived_time_series/
      - csv for 01-22-2020 to 03-23-2020
  - csse_covid_19_data/
    - csse_covid_19_daily_reports/
      - csv from 2020 to now (1/5/2022)
    - csse_covid_19_daily_reports_us/
      - csv from 2020 to now (1/5/2022)
    - csse_covid_19_time_series/
      - csv from 2020 to now (1/5/2022)
    - UID_ISO_FIPS_LookUp_Table.csv → have population data for country/province
  - who_covid_19_situation_reports/
    - who_covid_19_sit_rep_pdfs/ → pdf files
    - who_covid_19_sit_rep_time_series/ → 1 csv file for data in 2020 partially



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	FIPS	Admin2	Province_	Country_F	Last_Upda	Lat	Long_	Confirmec	Deaths	Recoverec	Active	Combined	Incident_F	Case_Fatality_Ratio	
2				Afghanista	########	33.93911	67.70995	53011	2237	42666	8108	Afghanista	136.1759	4.219879	
3				Albania	########	41.1533	20.1683	60283	1210	35551	23522	Albania	2094.76	2.007199	
4				Algeria	########	28.0339	1.6596	100873	2782	68185	29906	Algeria	230.0356	2.757923	
5				Andorra	########	42.5063	1.5218	8308	84	7585	639	Andorra	10752.6	1.011074	
6				Angola	########	-11.2027	17.8739	17756	410	11376	5970	Angola	54.025	2.309079	
7				Antigua ar	########	17.0608	-61.7964	160	5	148	7	Antigua ar	163.3853	3.125	
8				Argentina	########	-38.4161	-63.6167	1662730	43785	1466616	152329	Argentina	3678.95	2.63332	
9				Armenia	########	40.0691	45.0382	160544	2878	145759	11907	Armenia	5417.864	1.792655	
10			Australian	Australia	########	-35.4735	149.0124	118	3	115	0	Australian	27.56365	2.542373	
11			New Soutl	Australia	########	-33.8688	151.2093	4978	54	0	4924	New Soutl	61.32052	1.084773	
12			Northern `	Australia	########	-12.4634	130.8456	87	0	70	17	Northern `	35.42345	0	
13			Queenslar	Australia	########	-27.4698	153.0251	1263	6	1233	24	Queenslar	24.68967	0.475059	
14			South Aus	Australia	########	-34.9285	138.6007	582	4	567	11	South Aus	33.13407	0.687285	
15			Tasmania	Australia	########	-42.8821	147.3272	234	13	221	0	Tasmania,	43.69748	5.555556	
16			Victoria	Australia	########	-37.8136	144.9631	20402	820	19541	41	Victoria, A	307.7271	4.019214	
17			Western A	Australia	########	-31.9505	115.8605	872	9	845	18	Western A	33.14833	1.03211	
18				Austria	########	47.5162	14.5501	369721	6457	343039	20225	Austria	4105.092	1.746452	

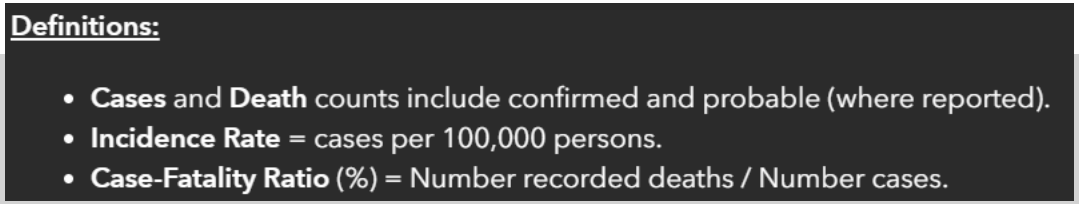


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Province_	Country_F	Last_Upda	Lat	Long_	Confirmec	Deaths	Recoverec	Active	FIPS	Incident_F	Total_Test	People_Hc	Case_Fata	UID	ISO3	Testing_R	Hospitalization_Rate	
2	Alabama	US	########	32.3182	-86.9023	389230	5080	211684	172466	1	7938.309	3360666		1.305141	84000001	USA	68540.47		
3	Alaska	US	########	61.3707	-152.404	49255	223	7165	41867	2	6733.01	1326542		0.452746	84000002	USA	181334.3		
4	American	US	########	-14.271	-170.132	0	0			60	0	2140			16	ASM	3846.085		
5	Arizona	US	########	33.7298	-111.431	584593	9741	82155	492697	4	8031.539	5416054		1.666287	84000004	USA	41076.09		
6	Arkansas	US	########	34.9697	-92.3731	245916	3926	215980	26010	5	8148.839	2129230		1.59648	84000005	USA	70555.61		
7	California	US	########	36.1162	-119.682	2676461	28313			6	6774.157	34760355		1.057752	84000006	USA	87973.68		
8	Colorado	US	########	39.0598	-105.311	352923	5102	18775	329046	8	6128.48	4649434		1.445641	84000008	USA	80737.06		
9	Connectic	US	########	41.5978	-72.7554	202758	6287	9800	186671	9	5687.004	4581696		3.100741	84000009	USA	128508.5		
10	Delaware	US	########	39.3185	-75.5071	62949	1112	18851	42986	10	6464.503	1029436		1.766509	84000010	USA	105717.2		
11	Diamond I	US	########			49	0			88888				0	84088888	USA			
12	District of	US	########	38.8974	-77.0268	30750	808	21660	8282	11	4357.073	937035		2.627642	84000011	USA	132771.7		
13	Florida	US	########	27.7663	-81.6868	1429722	22481			12	6656.763	18070454		1.572404	84000012	USA	84135.74		
14	Georgia	US	########	33.0406	-83.6431	723630	11230			13	6815.496	5569613		1.551898	84000013	USA	52457.3		
15	Grand Prir	US	########			103	3			99999				2.912621	84099999	USA			
16	Guam	US	########	13.4443	144.7937	7386	124	7145	117	66	4497.379	98270		1.678852	316	GUM	59837.18		
17	Hawaii	US	########	21.0943	-157.498	23400	299	11958	11143	15	1652.692	849892		1.277778	84000015	USA	60026.05		
18	Idaho	US	########	44.2405	-114.479	147173	1517	63097	82559	16	8235.459	554296		1.03076	84000016	USA	31017.11		
19	Illinois	US	########	40.3495	-88.9861	1008045	18941			17	7955.013	13803946		1.878984	84000017	USA	108934.2		
20	Indiana	US	########	39.8494	-86.2583	546499	9958	360087	176454	18	8117.665	5950922		1.822144	84000018	USA	88394.66		
21	Iowa	US	########	42.0115	-93.2105	293189	4065	251659	37465	19	9292.631	1188908		1.386478	84000019	USA	37682.46		



	A	B	C	D	E	F	G	H
1	Province/S	Country/R	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20
2		Afghanista	33.93911	67.70995	0	0	0	0
3		Albania	41.1533	20.1683	0	0	0	0
4		Algeria	28.0339	1.6596	0	0	0	0
5		Andorra	42.5063	1.5218	0	0	0	0
6		Angola	-11.2027	17.8739	0	0	0	0
7		Antarctica	-71.9499	23.347	0	0	0	0
8		Antigua ar	17.0608	-61.7964	0	0	0	0
9		Argentina	-38.4161	-63.6167	0	0	0	0

CSSE Covid 19 Daily Report Schema



Definitions:

- **Cases** and **Death** counts include confirmed and probable (where reported).
- **Incidence Rate** = cases per 100,000 persons.
- **Case-Fatality Ratio** (%) = Number recorded deaths / Number cases.

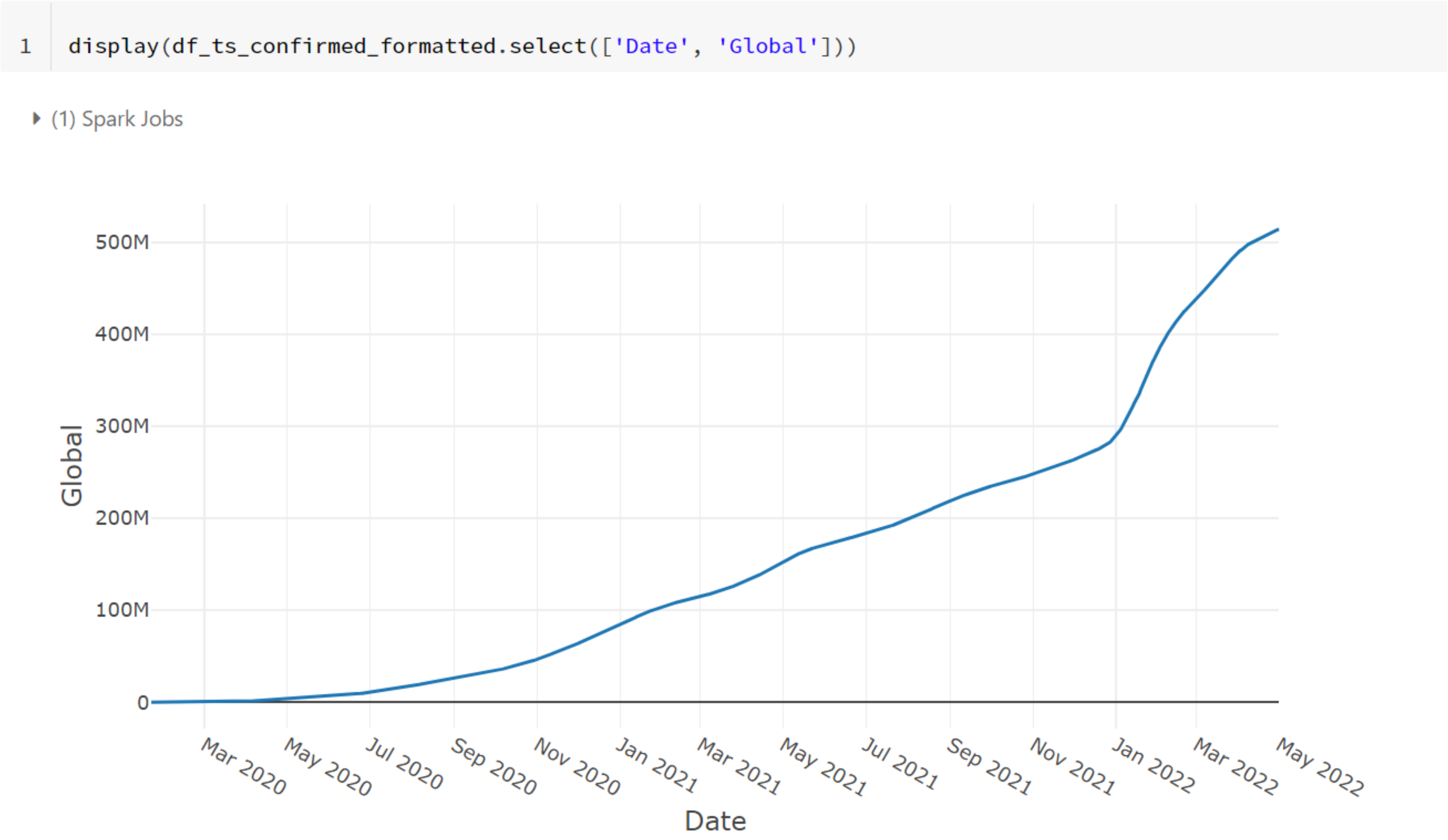
- Incidence Rate
  - The number of new cases for the disease within a time frame, as a proportion of the number of people at risk for the disease
  - Allows for easy comparison of the cases across geographies with different populations – specifically as total COVID-19 cases for every 100,000 people
  - Provides a complete picture of the state of the pandemic in a community
  - Reference: https://www.usf.edu/business/state-of-the-region/e-insights-2021/section-2-01-covid-19-incidence-rate.aspx
- Case Fatality Ratio
  - Proportion of individuals diagnosed with a disease who die from that disease and is therefore a measure of severity among detected cases
  - Reference: https://www.who.int/news-room/commentaries/detail/estimating-mortality-from-covid-19#:~:text=Calculating%20CFR%20Case%20fatality%20ratio,severity%20among%20detected%20cases%3A

Global Overall COVID-19 Situation

- Globally (as of 1/5/2022):
  - Total number of confirmed COVID-19 cases is around 514M (28-Day: around 21.5M)
  - Total number of death cases is around 6M (28-Day: around 81K)
  - Incidence Rate is at around 22K
    - i.e., There are a total of around 22K new COVID-19 cases for every 100,000 people in the world
  - Case Fatality Rate is at around 2.29
    - i.e., There are around 2.29% of individuals diagnosed COVID-19 died as of 1st May 2022.

Objectives	Figures
Total Confirmed Cases	513,869,166
Total Deaths	6,236,553
Incidence Rate (for every 100,000 population)	22549.5129
Case Fatality Rate	2.2867%
28-Day Total Confirmed Cases	21,550,658
28-Day Total Death Cases	81,164

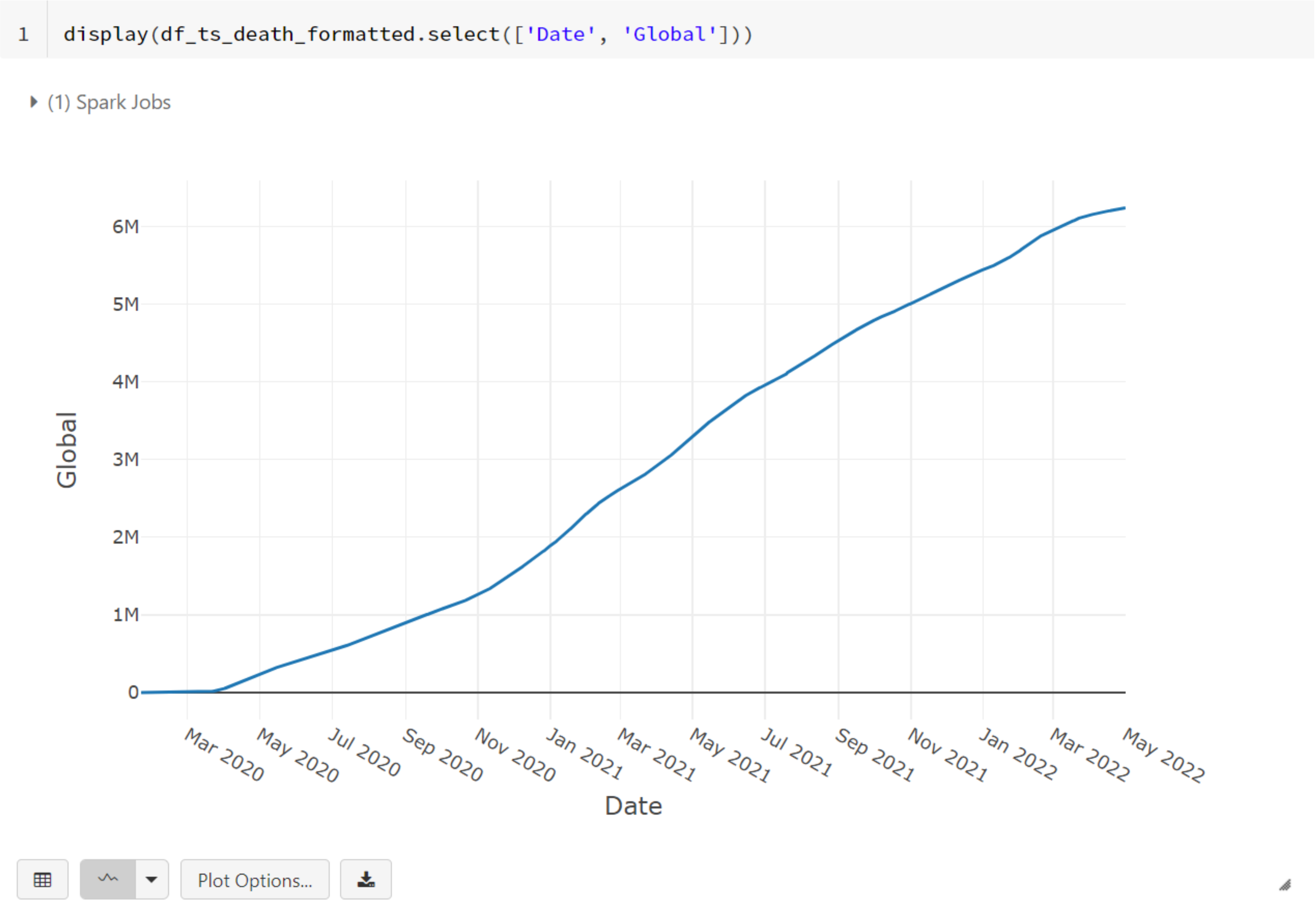
Global Cumulative Confirmed COVID-19 Cases



```
display(df_ts_confirmed_formatted.select(['Date', 'Global']))
```

(1) Spark Jobs

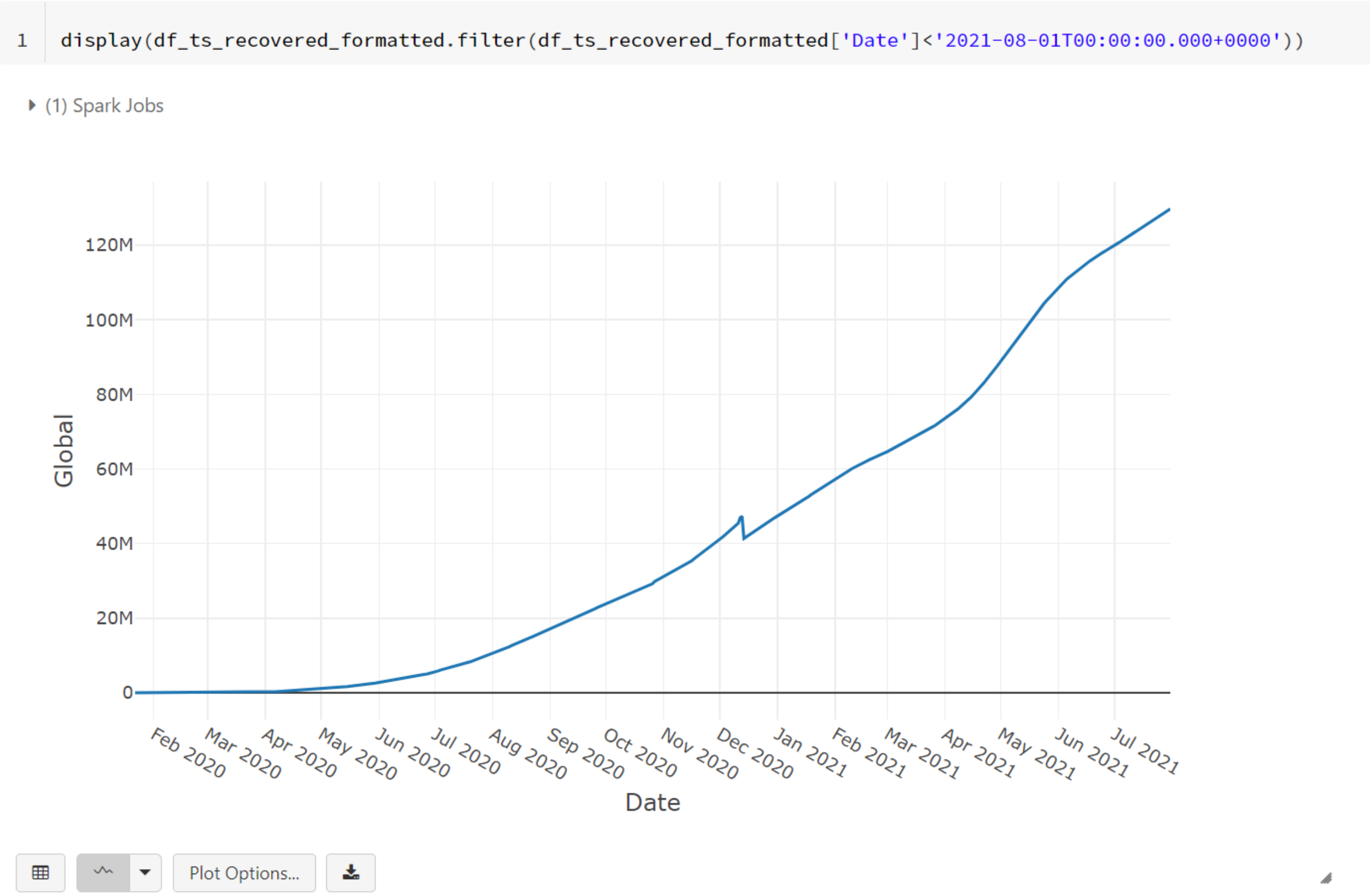
Global Cumulative COVID-19 Death Cases



```
display(df_ts_death_formatted.select(['Date', 'Global']))
```

Global Cumulative COVID-19 Recovered Cases

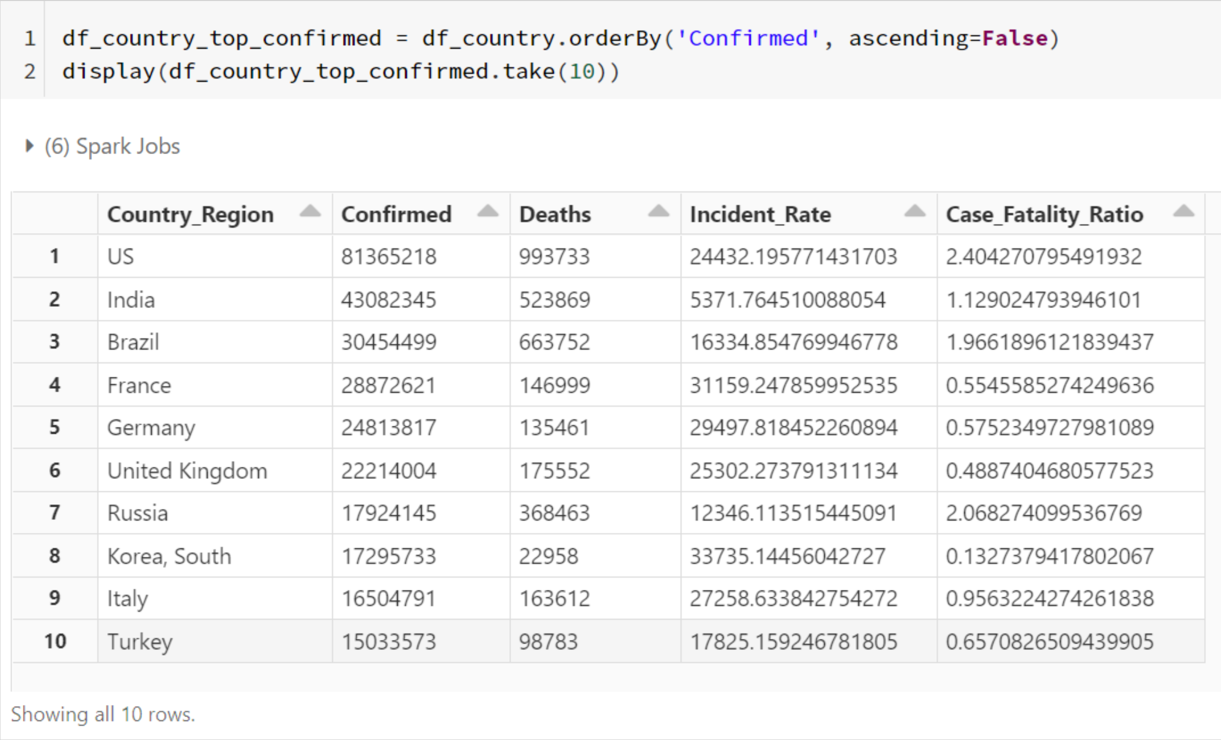
Note: Data after August 2021 not available



```
display(df_ts_recovered_formatted.filter(df_ts_recovered_formatted['Date']<'2021-08-01T00:00:00.000+0000'))
```

COVID-19 Situation By Countries

- Countries with Top 10 Confirmed COVID-19 Cases:
  - Insights: All of them are economically strong

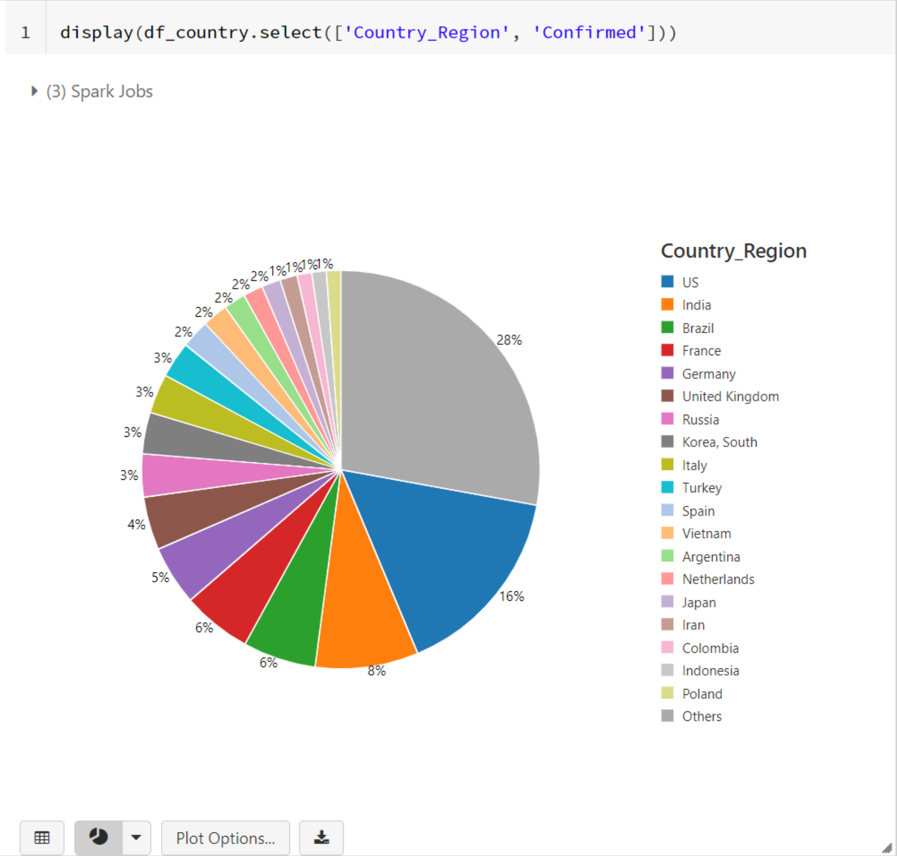


```
df_country_top_confirmed = df_country.orderBy('Confirmed', ascending=False)
display(df_country_top_confirmed.take(10))
```

▸ (6) Spark Jobs

	Country_Region	Confirmed	Deaths	Incident_Rate	Case_Fatality_Ratio
1	US	81365218	993733	24432.195771431703	2.404270795491932
2	India	43082345	523869	5371.764510088054	1.129024793946101
3	Brazil	30454499	663752	16334.854769946778	1.9661896121839437
4	France	28872621	146999	31159.247859952535	0.5545585274249636
5	Germany	24813817	135461	29497.818452260894	0.5752349727981089
6	United Kingdom	22214004	175552	25302.273791311134	0.4887404680577523
7	Russia	17924145	368463	12346.113515445091	2.068274099536769
8	Korea, South	17295733	22958	33735.14456042727	0.1327379417802067
9	Italy	16504791	163612	27258.633842754272	0.9563224274261838
10	Turkey	15033573	98783	17825.159246781805	0.6570826509439905

Showing all 10 rows.

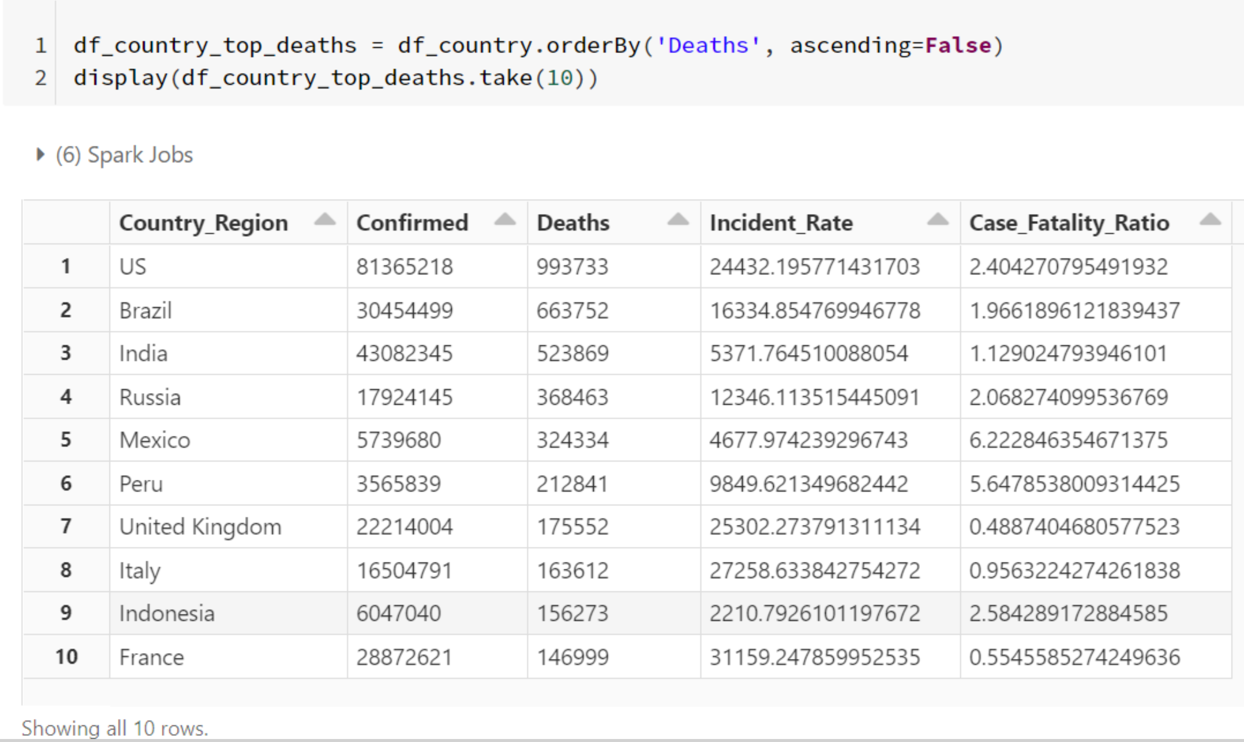


```
display(df_country.select(['Country_Region', 'Confirmed']))
```

▸ (3) Spark Jobs

COVID-19 Situation By Countries

- Countries with Top 10 COVID-19 Deaths:

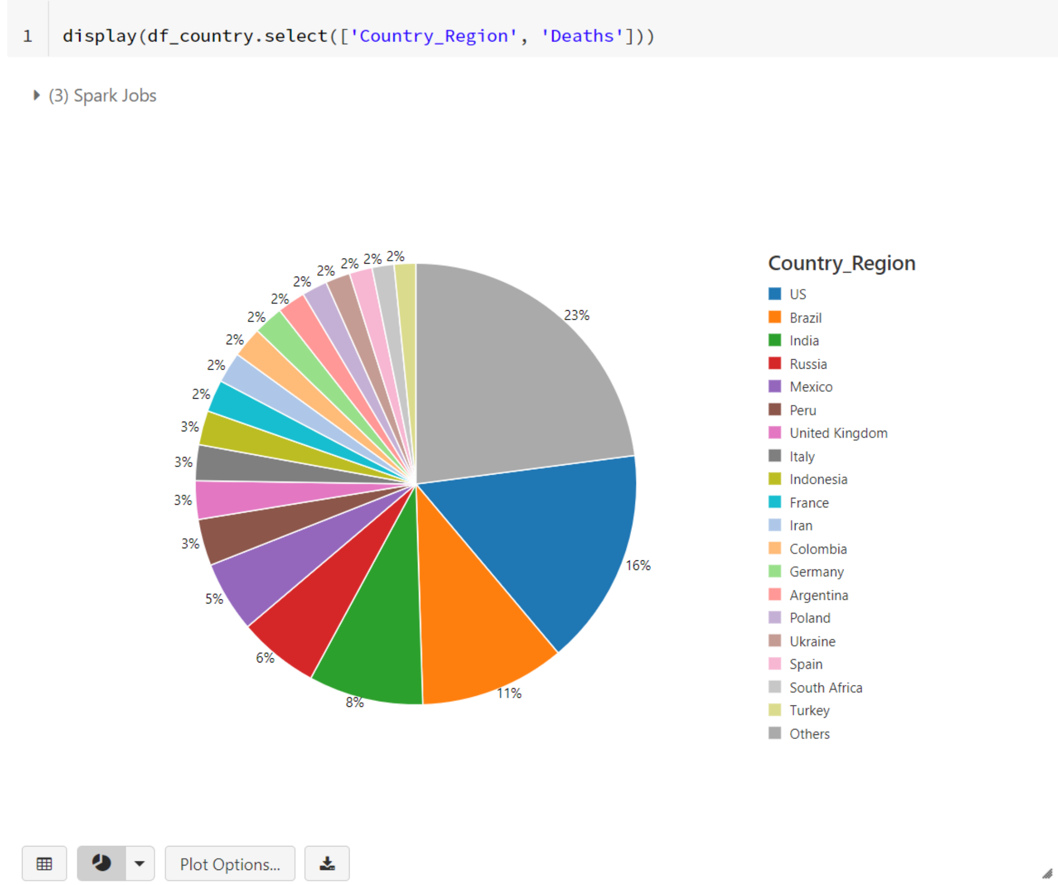


```
df_country_top_deaths = df_country.orderBy('Deaths', ascending=False)
display(df_country_top_deaths.take(10))
```

▸ (6) Spark Jobs

	Country_Region	Confirmed	Deaths	Incident_Rate	Case_Fatality_Ratio
1	US	81365218	993733	24432.195771431703	2.404270795491932
2	Brazil	30454499	663752	16334.854769946778	1.9661896121839437
3	India	43082345	523869	5371.764510088054	1.129024793946101
4	Russia	17924145	368463	12346.113515445091	2.068274099536769
5	Mexico	5739680	324334	4677.974239296743	6.222846354671375
6	Peru	3565839	212841	9849.621349682442	5.6478538009314425
7	United Kingdom	22214004	175552	25302.273791311134	0.4887404680577523
8	Italy	16504791	163612	27258.633842754272	0.9563224274261838
9	Indonesia	6047040	156273	2210.7926101197672	2.584289172884585
10	France	28872621	146999	31159.247859952535	0.5545585274249636

Showing all 10 rows.

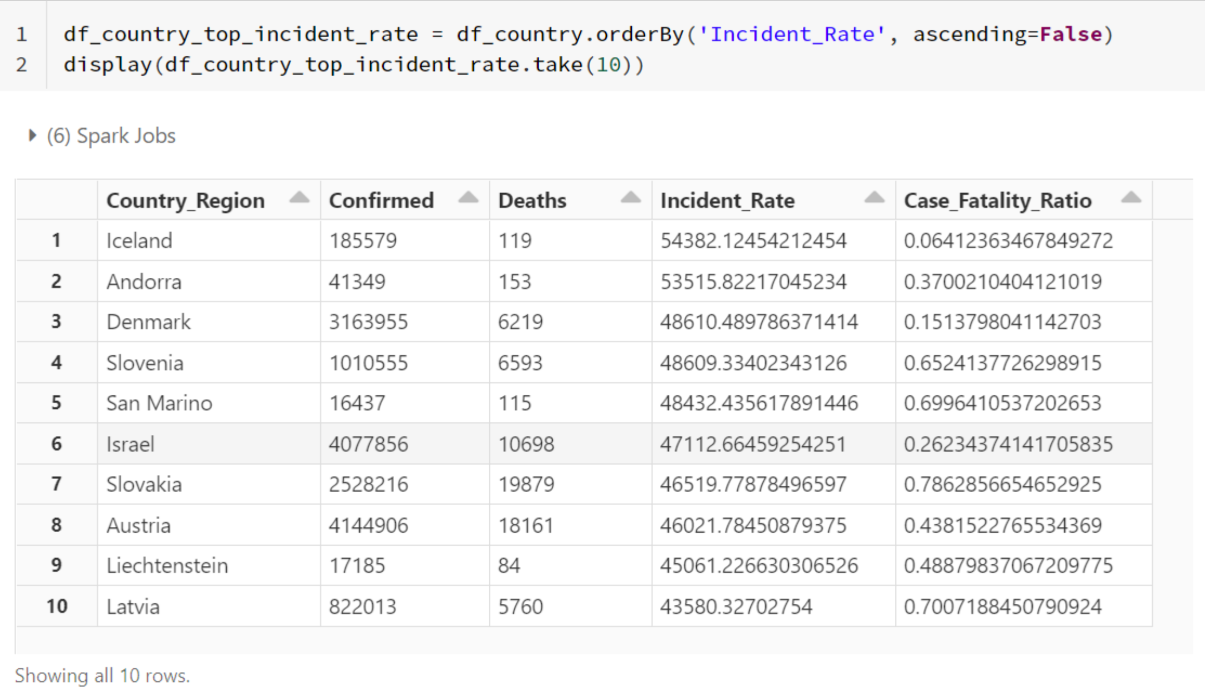


```
display(df_country.select(['Country_Region', 'Deaths']))
```

▸ (3) Spark Jobs

COVID-19 Situation By Countries

- Countries with Top 10 Incidence Rate:

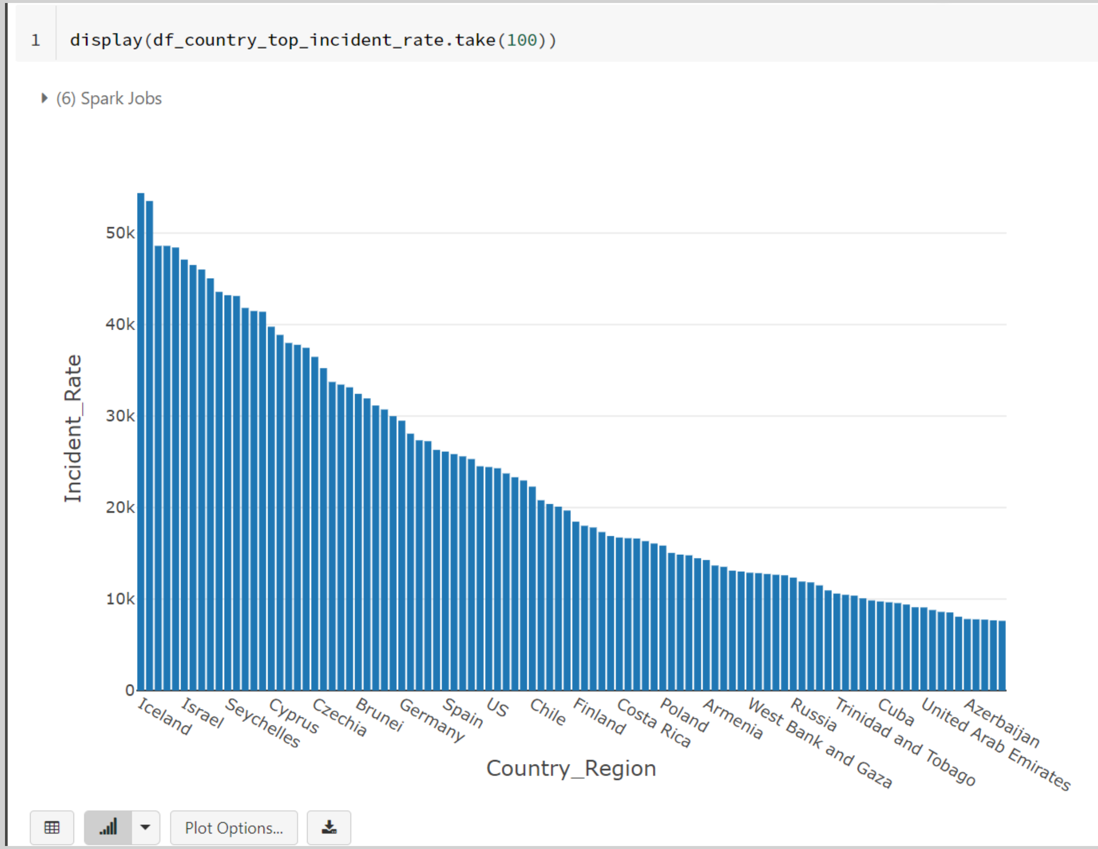


```
df_country_top_incident_rate = df_country.orderBy('Incident_Rate', ascending=False)
display(df_country_top_incident_rate.take(10))
```

(6) Spark Jobs

	Country_Region	Confirmed	Deaths	Incident_Rate	Case_Fatality_Ratio
1	Iceland	185579	119	54382.12454212454	0.06412363467849272
2	Andorra	41349	153	53515.82217045234	0.3700210404121019
3	Denmark	3163955	6219	48610.489786371414	0.1513798041142703
4	Slovenia	1010555	6593	48609.33402343126	0.6524137726298915
5	San Marino	16437	115	48432.435617891446	0.6996410537202653
6	Israel	4077856	10698	47112.66459254251	0.26234374141705835
7	Slovakia	2528216	19879	46519.77878496597	0.7862856654652925
8	Austria	4144906	18161	46021.78450879375	0.4381522765534369
9	Liechtenstein	17185	84	45061.226630306526	0.48879837067209775
10	Latvia	822013	5760	43580.32702754	0.7007188450790924

Showing all 10 rows.

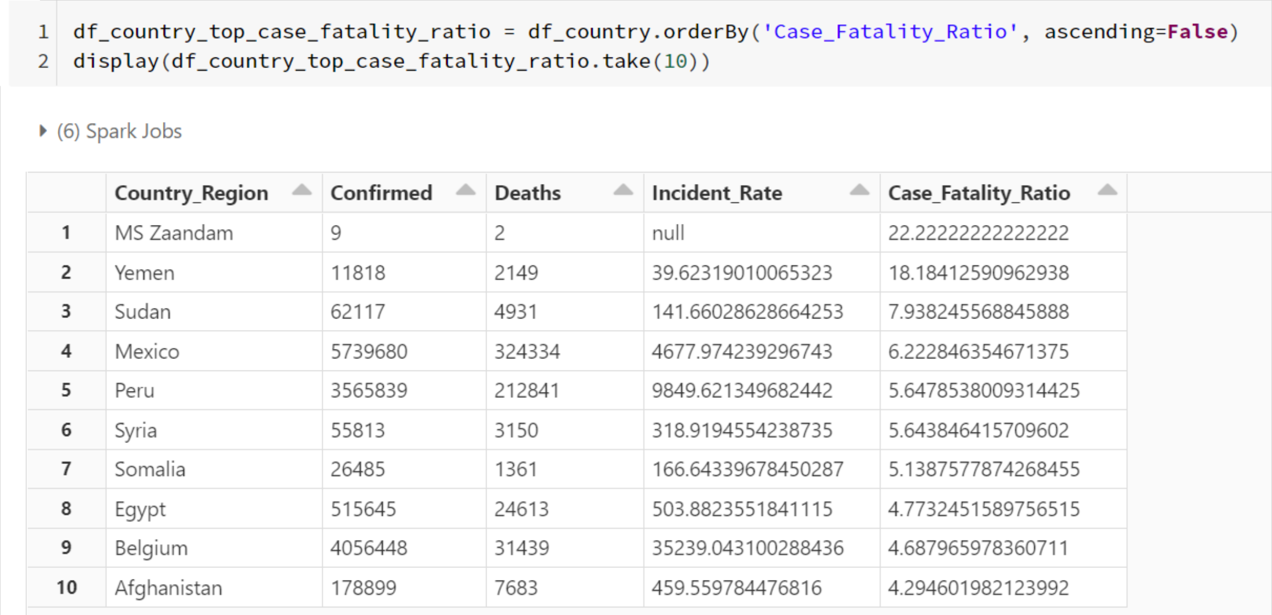


```
display(df_country_top_incident_rate.take(100))
```

(6) Spark Jobs

COVID-19 Situation By Countries

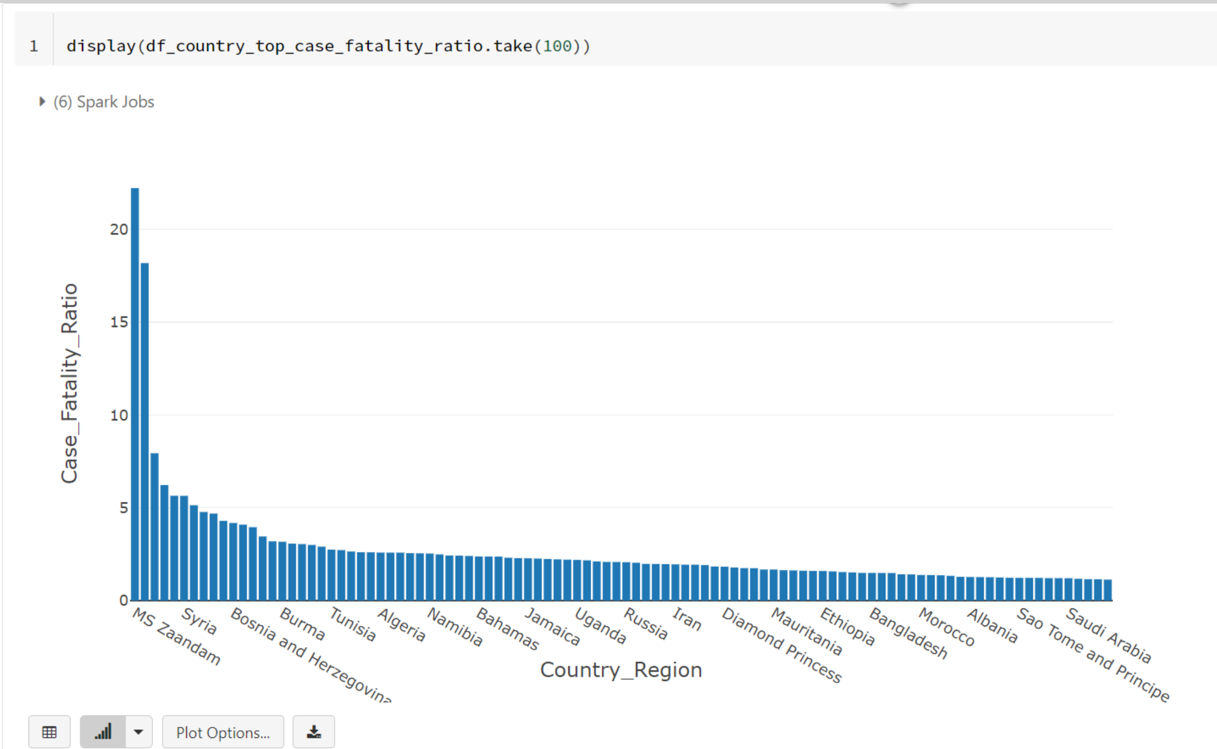
- Countries with Top 10 Case Fatality Rate:



```
df_country_top_case_fatality_ratio = df_country.orderBy('Case_Fatality_Ratio', ascending=False)
display(df_country_top_case_fatality_ratio.take(10))
```

(6) Spark Jobs

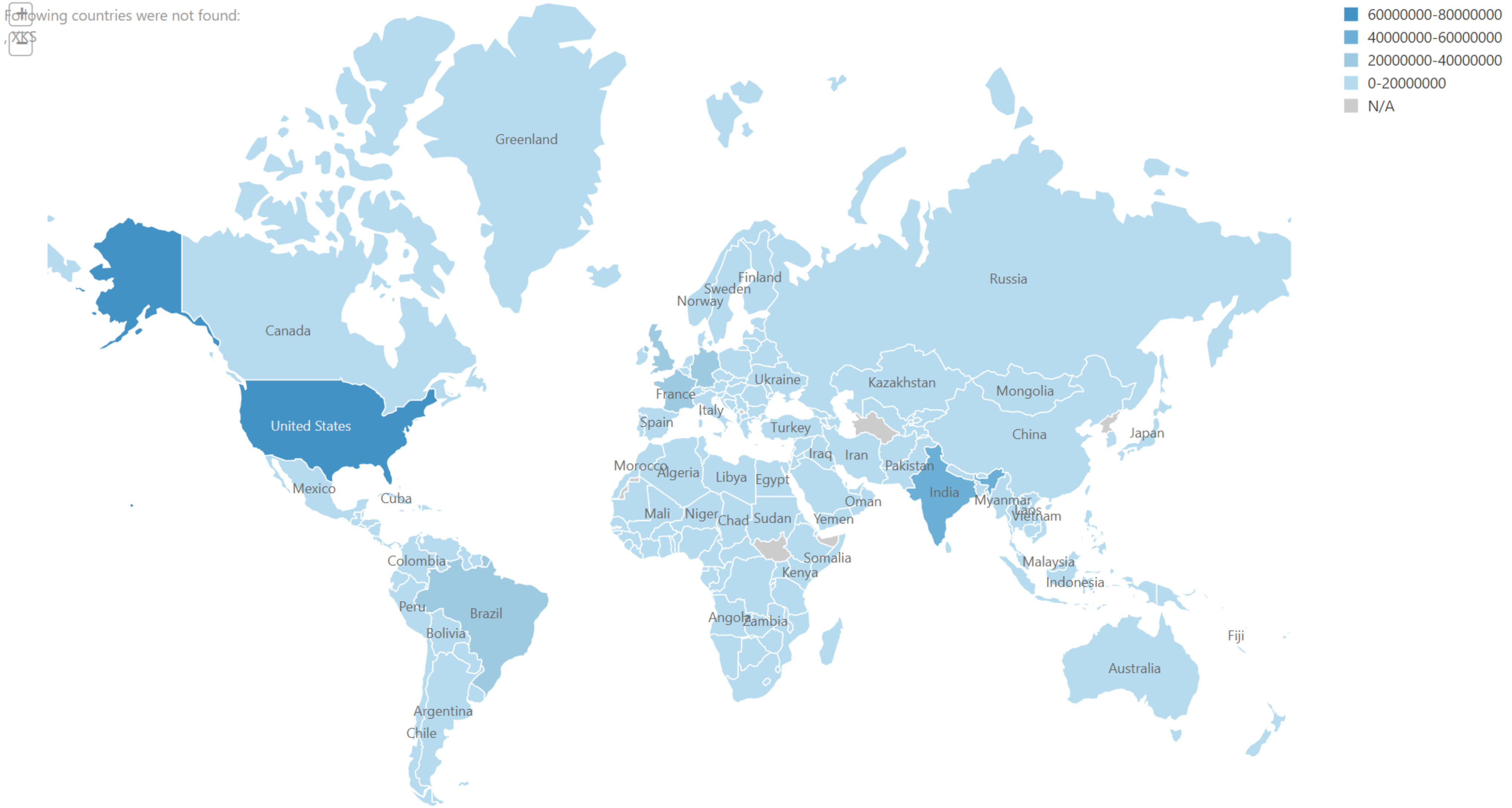
	Country_Region	Confirmed	Deaths	Incident_Rate	Case_Fatality_Ratio
1	MS Zaandam	9	2	null	22.22222222222222
2	Yemen	11818	2149	39.62319010065323	18.18412590962938
3	Sudan	62117	4931	141.66028628664253	7.938245568845888
4	Mexico	5739680	324334	4677.974239296743	6.222846354671375
5	Peru	3565839	212841	9849.621349682442	5.6478538009314425
6	Syria	55813	3150	318.9194554238735	5.643846415709602
7	Somalia	26485	1361	166.64339678450287	5.1387577874268455
8	Egypt	515645	24613	503.8823551841115	4.7732451589756515
9	Belgium	4056448	31439	35239.043100288436	4.687965978360711
10	Afghanistan	178899	7683	459.559784476816	4.294601982123992



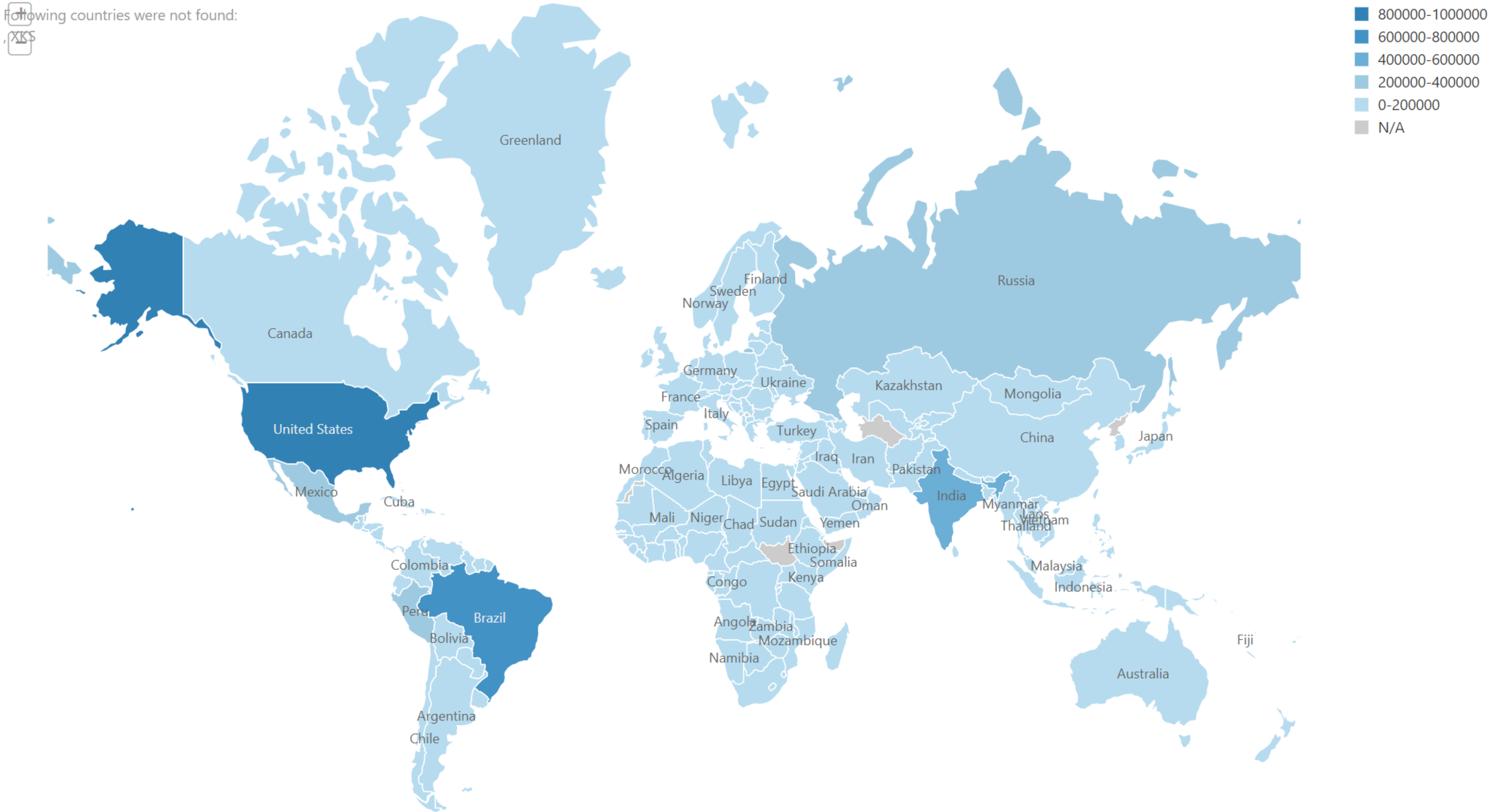
```
display(df_country_top_case_fatality_ratio.take(100))
```

(6) Spark Jobs

Global Distribution of Confirmed COVID-19 Cases

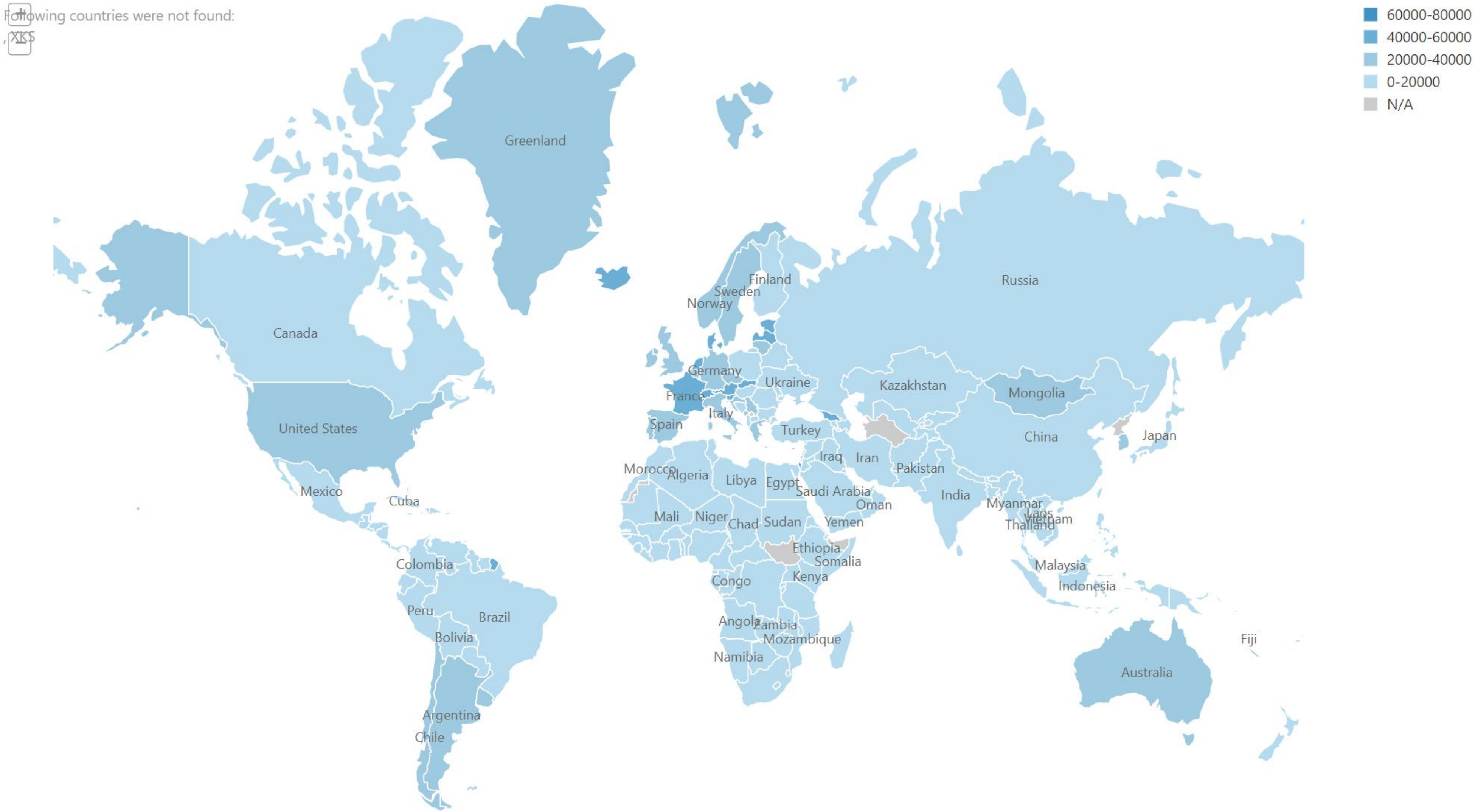


Global Distribution of COVID-19 Death Cases



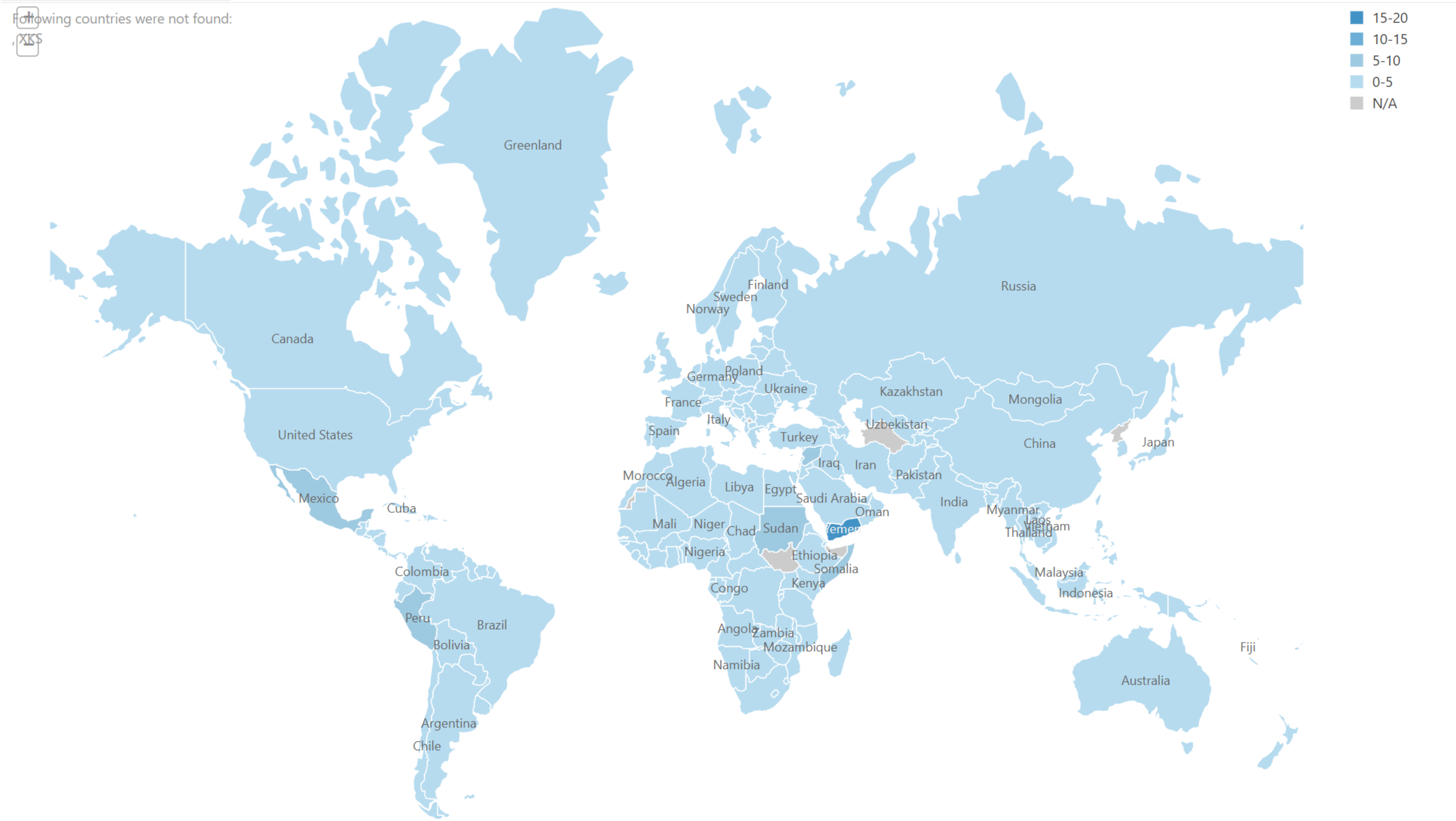
Following countries were not found:
,XKS

Global Distribution of COVID-19 Incidence Rate



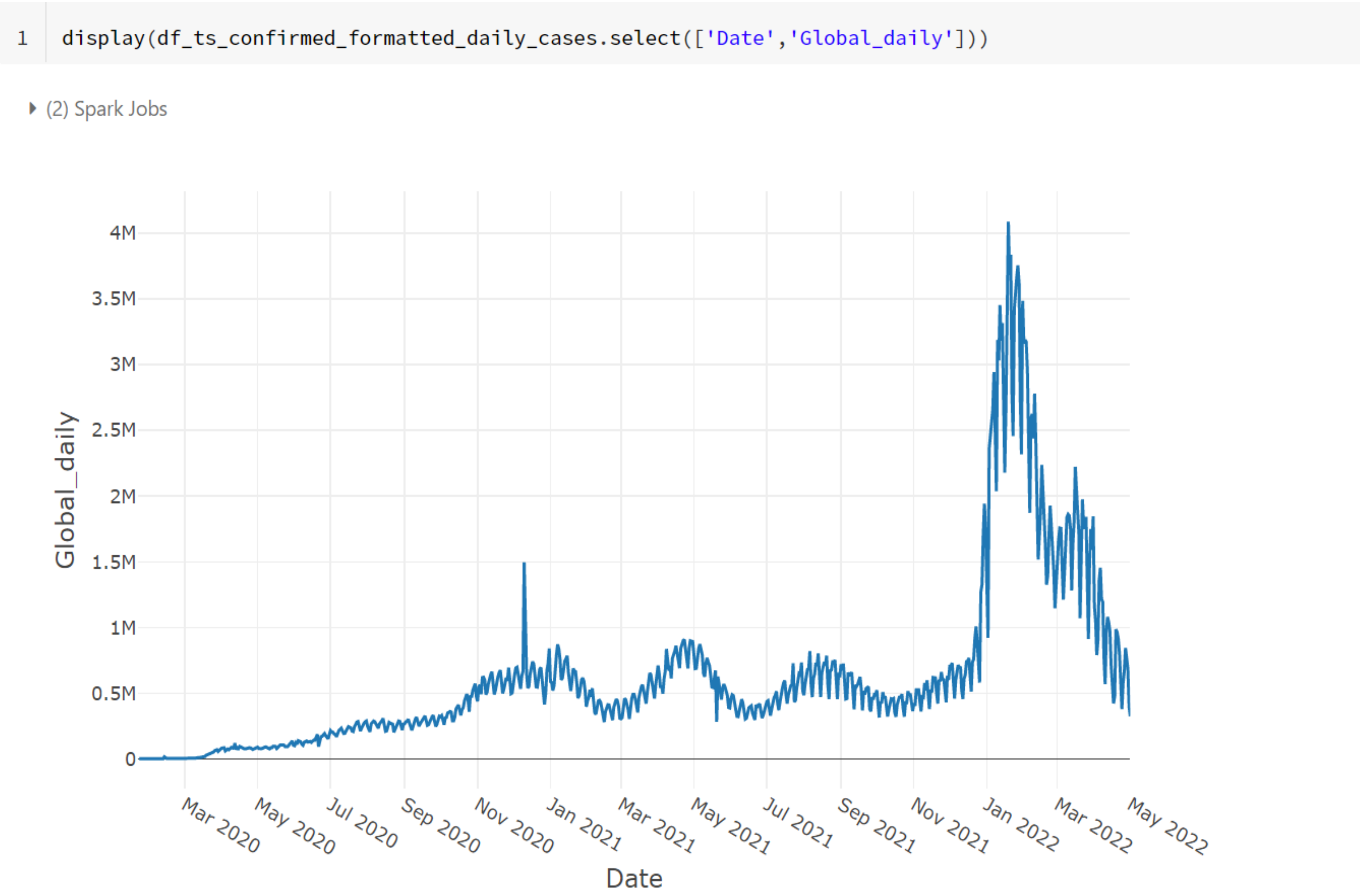
Following countries were not found:
,XKS

Global Distribution of COVID-19 Case Fatality Rate



Following countries were not found:
, XKS

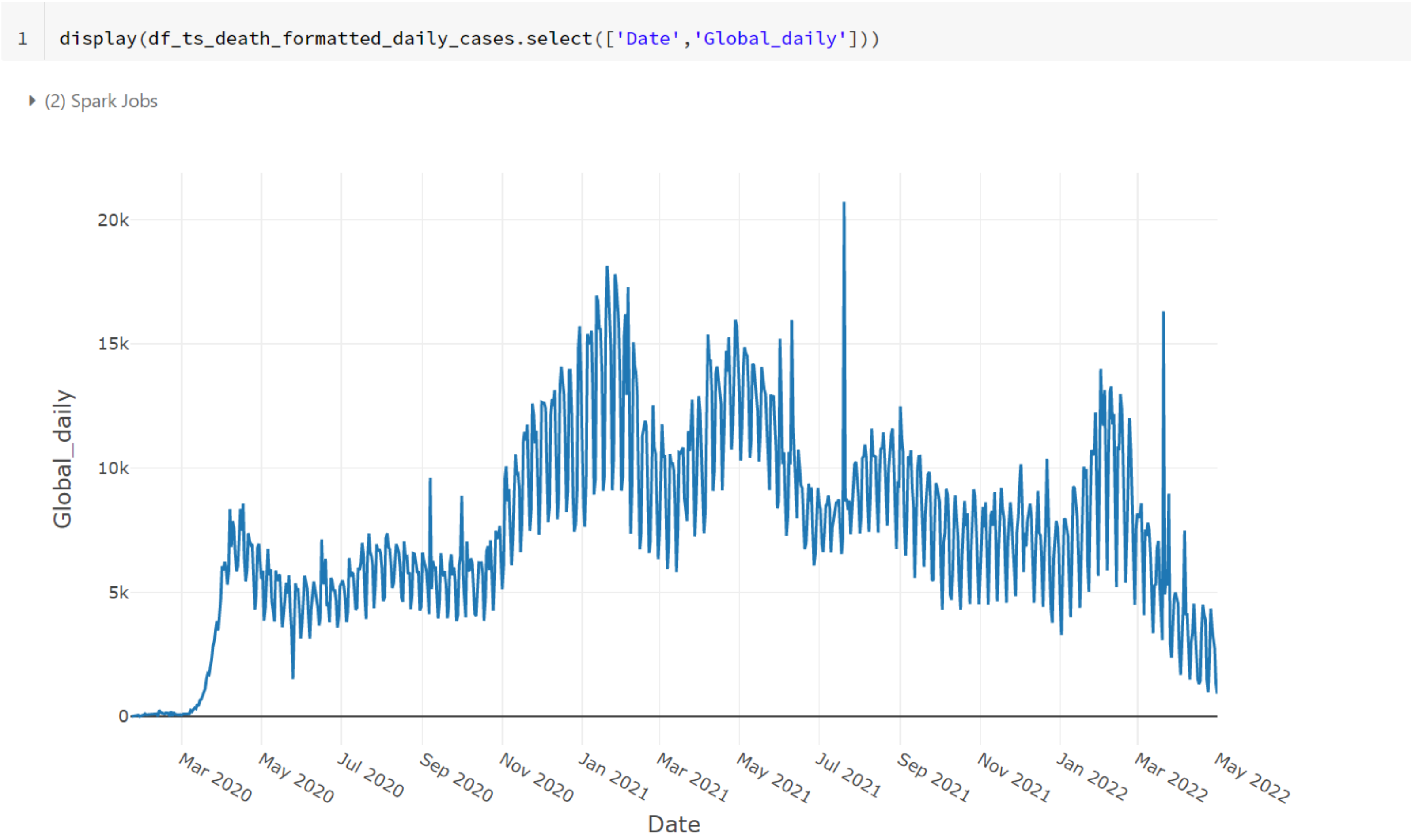
Global Daily Confirmed Cases



```
display(df_ts_confirmed_formatted_daily_cases.select(['Date','Global_daily']))
```

(2) Spark Jobs

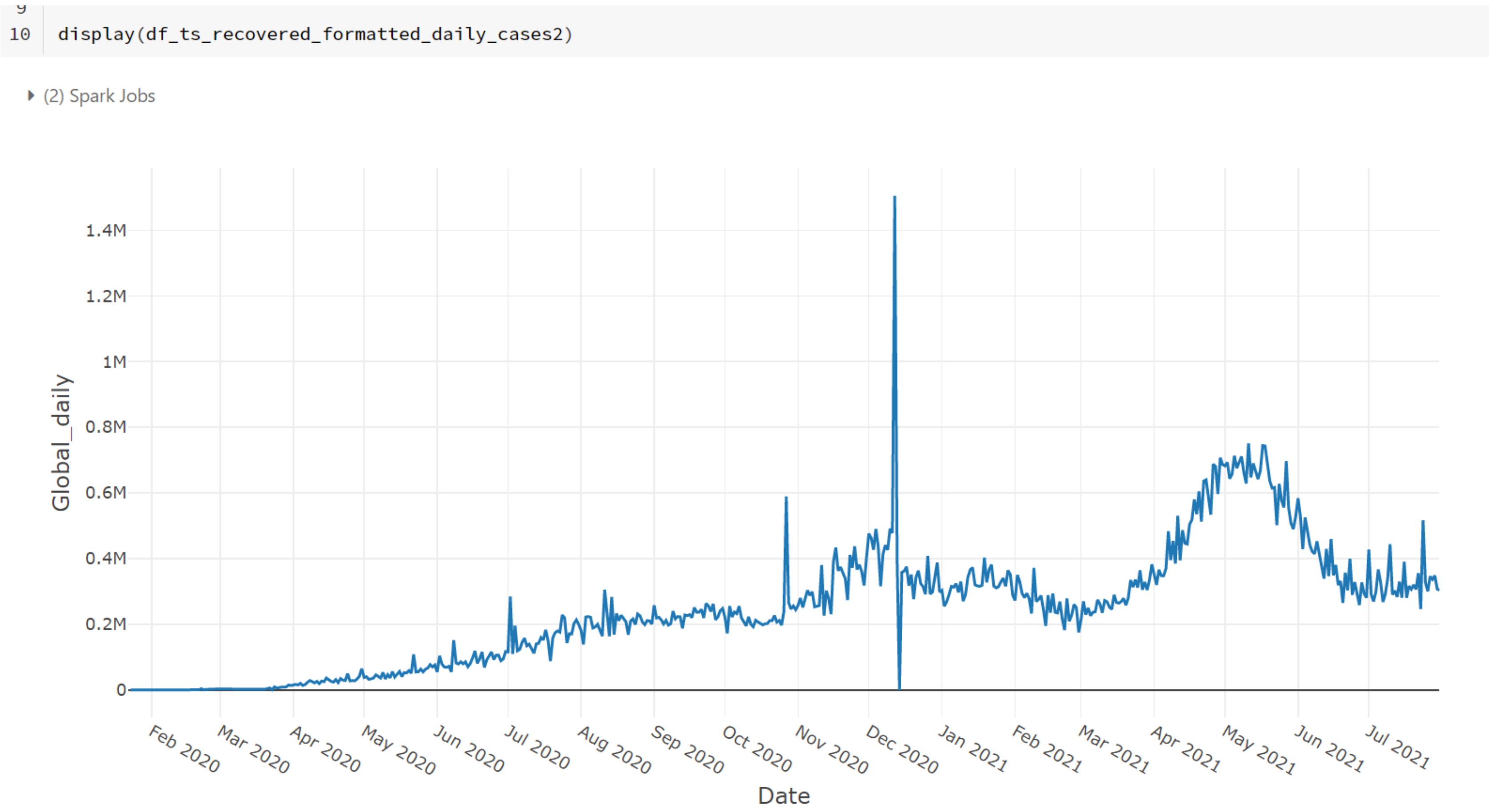
Global Daily Death Cases



```
display(df_ts_death_formatted_daily_cases.select(['Date','Global_daily']))
```

▸ (2) Spark Jobs

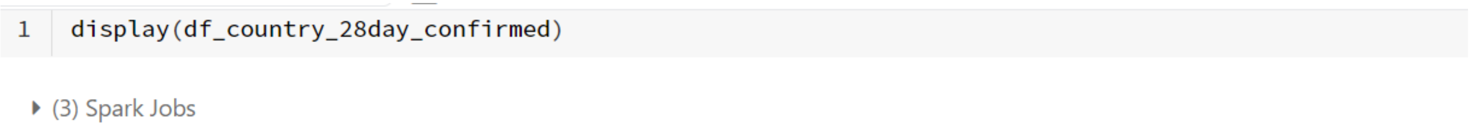
Global Daily Recovered Cases



```
display(df_ts_recovered_formatted_daily_cases2)
```

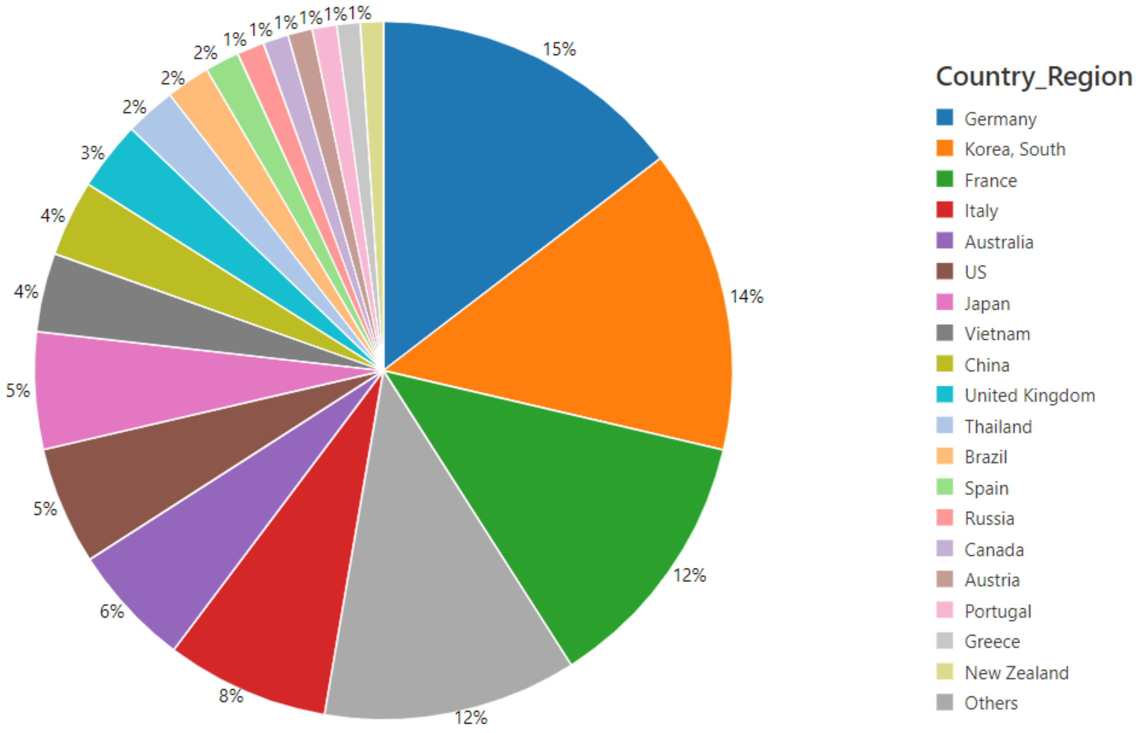
(2) Spark Jobs

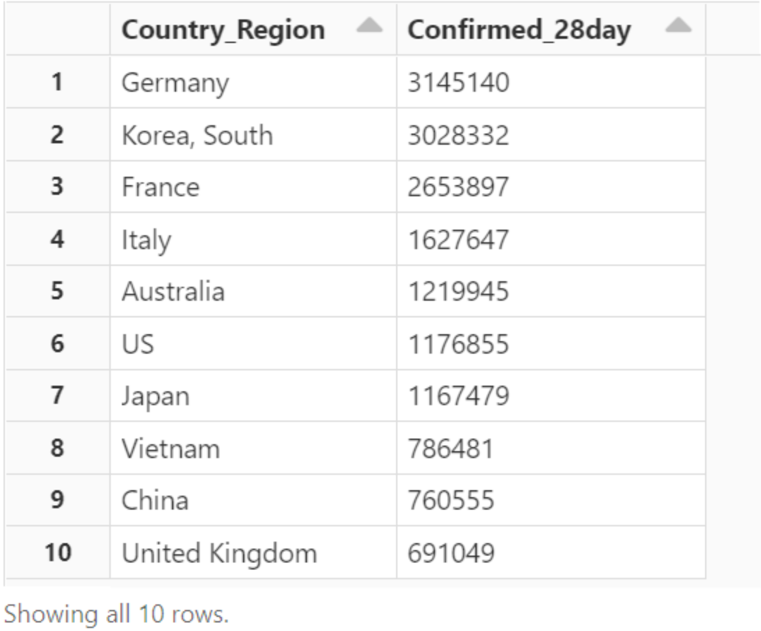
Top Countries with Highest 28-Day Confirmed COVID-19 Cases



```
display(df_country_28day_confirmed)
```

(3) Spark Jobs

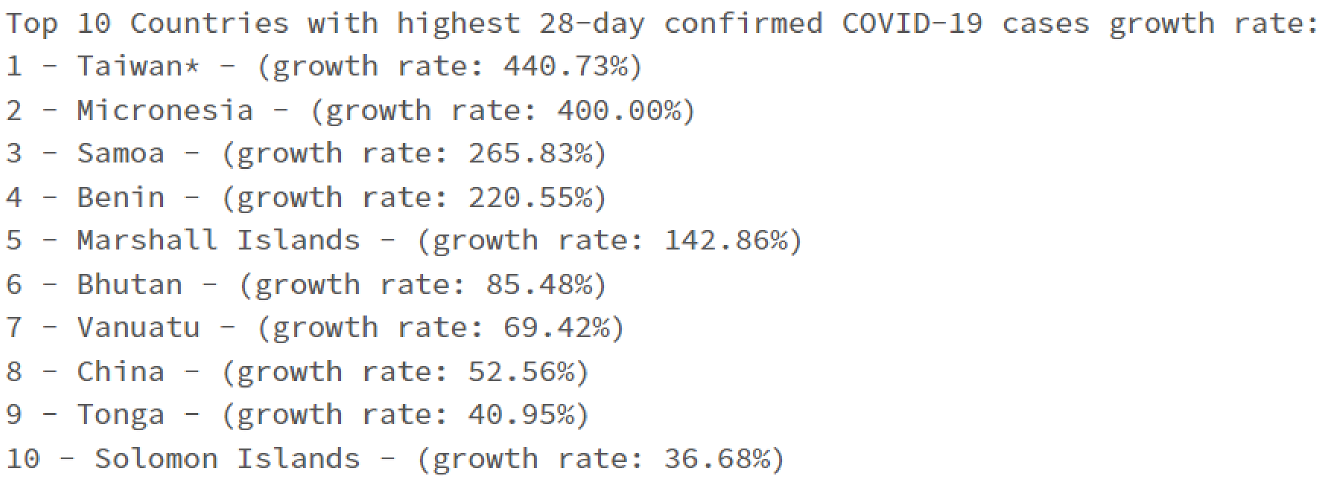




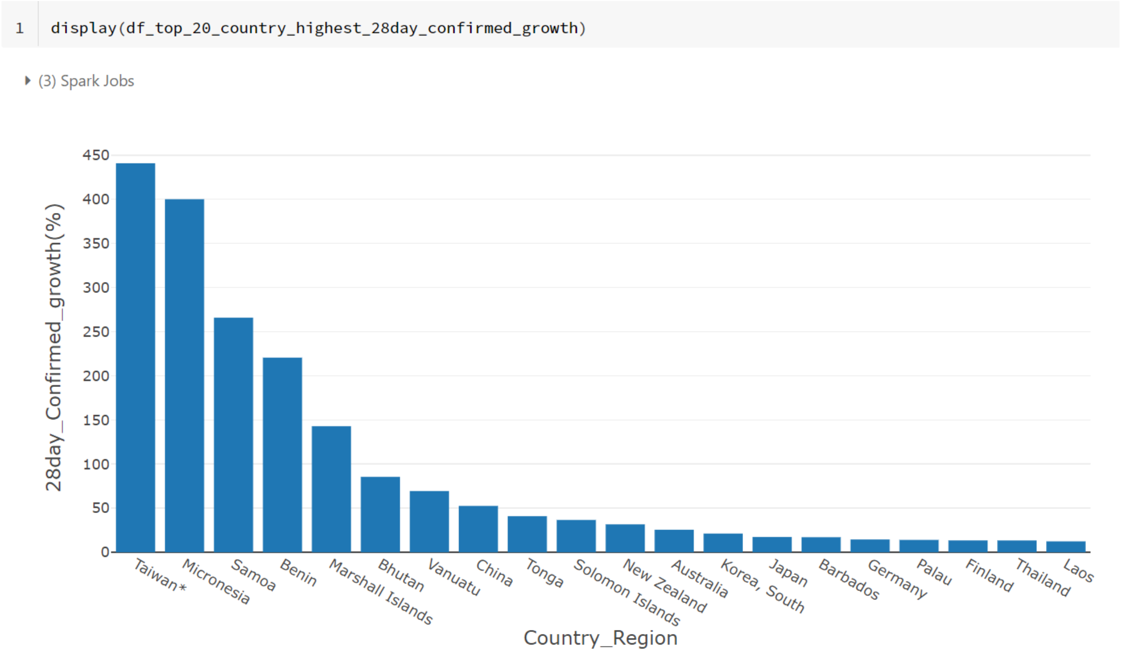
	Country_Region	Confirmed_28day
1	Germany	3145140
2	Korea, South	3028332
3	France	2653897
4	Italy	1627647
5	Australia	1219945
6	US	1176855
7	Japan	1167479
8	Vietnam	786481
9	China	760555
10	United Kingdom	691049

Showing all 10 rows.

Top Countries with Highest 28-Day Confirmed COVID-19 Cases Growth Rate

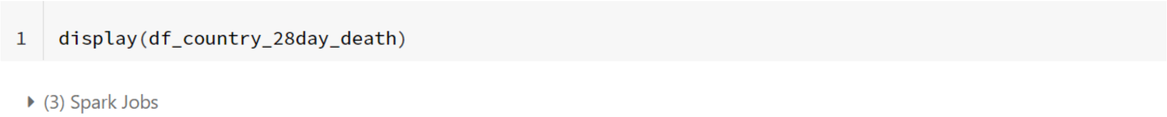


```
Top 10 Countries with highest 28-day confirmed COVID-19 cases growth rate:
1 - Taiwan* - (growth rate: 440.73%)
2 - Micronesia - (growth rate: 400.00%)
3 - Samoa - (growth rate: 265.83%)
4 - Benin - (growth rate: 220.55%)
5 - Marshall Islands - (growth rate: 142.86%)
6 - Bhutan - (growth rate: 85.48%)
7 - Vanuatu - (growth rate: 69.42%)
8 - China - (growth rate: 52.56%)
9 - Tonga - (growth rate: 40.95%)
10 - Solomon Islands - (growth rate: 36.68%)
```



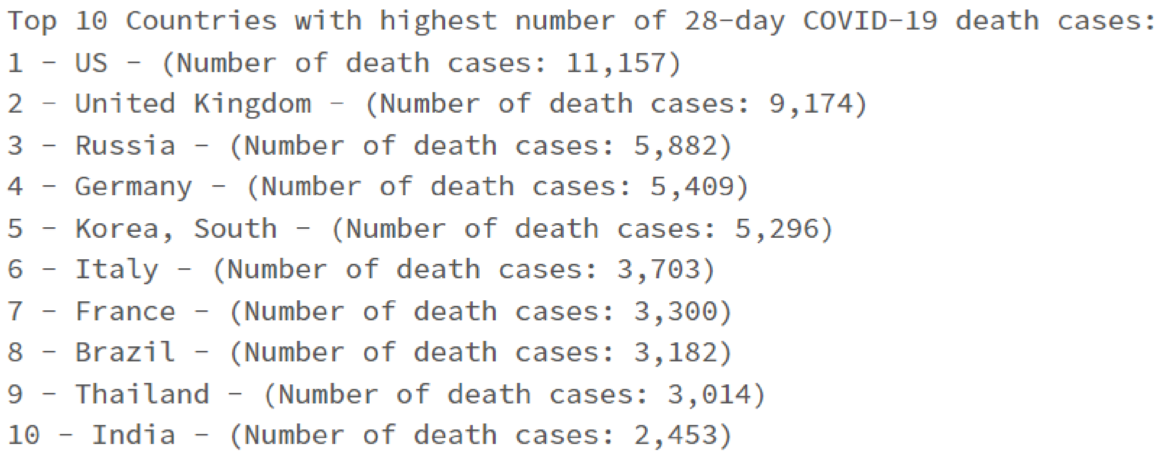
```
display(df_top_20_country_highest_28day_confirmed_growth)
```

Top Countries with Highest 28-Day COVID-19 Death Cases

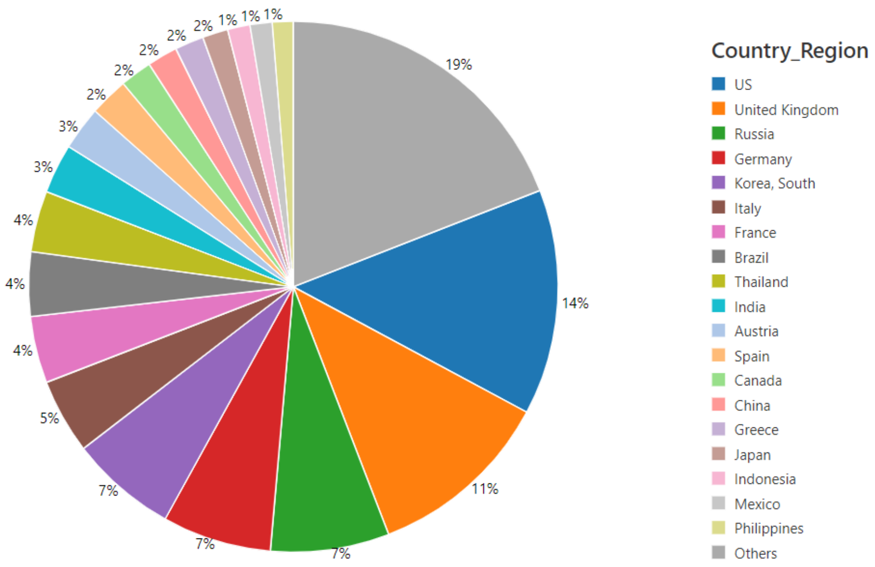


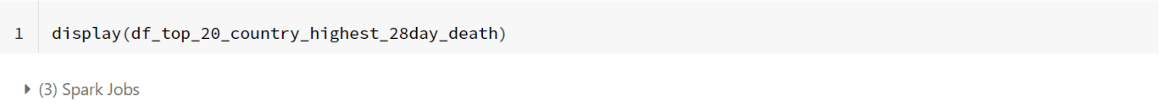
```
display(df_country_28day_death)
```

(3) Spark Jobs



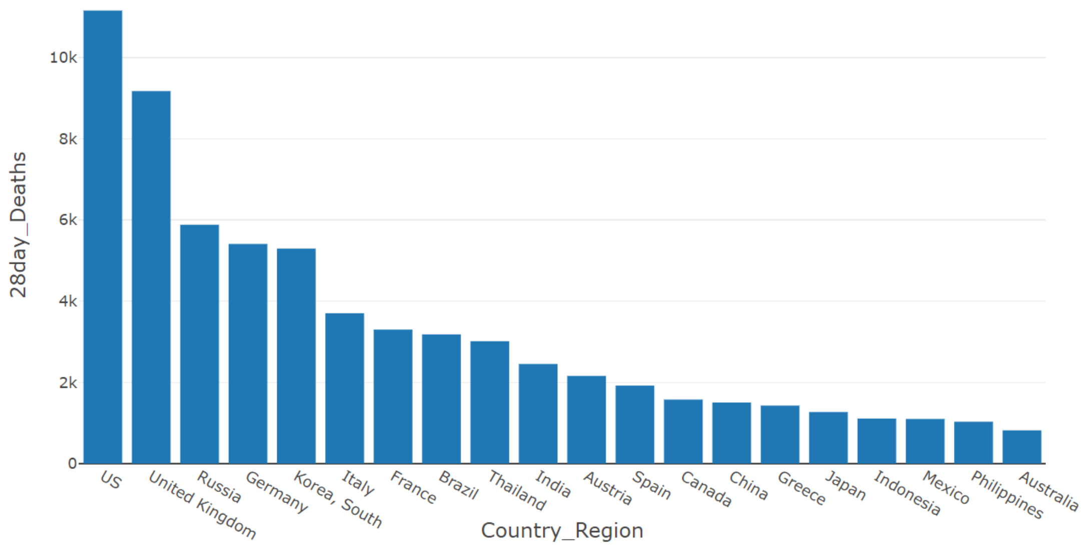
```
Top 10 Countries with highest number of 28-day COVID-19 death cases:
1 - US - (Number of death cases: 11,157)
2 - United Kingdom - (Number of death cases: 9,174)
3 - Russia - (Number of death cases: 5,882)
4 - Germany - (Number of death cases: 5,409)
5 - Korea, South - (Number of death cases: 5,296)
6 - Italy - (Number of death cases: 3,703)
7 - France - (Number of death cases: 3,300)
8 - Brazil - (Number of death cases: 3,182)
9 - Thailand - (Number of death cases: 3,014)
10 - India - (Number of death cases: 2,453)
```





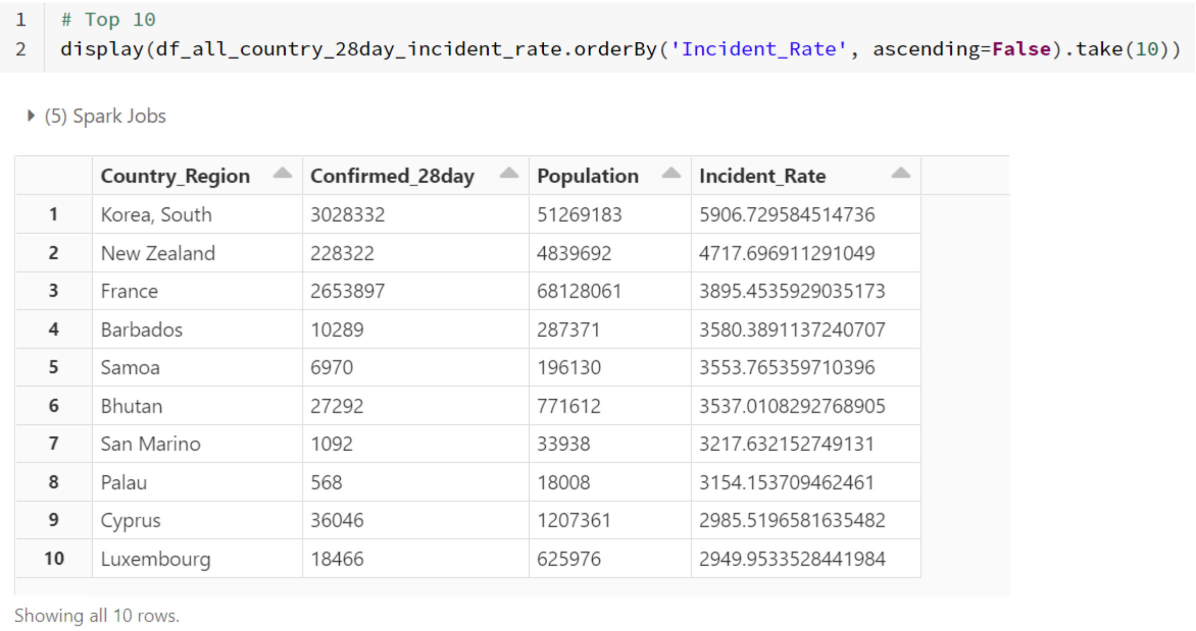
```
display(df_top_20_country_highest_28day_death)
```

(3) Spark Jobs



Top Countries with Highest 28-Day COVID-19 Incidence Rate

Note: for every 100,000 population

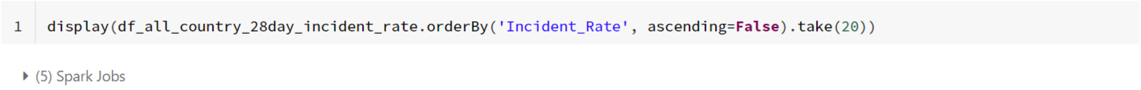


```
# Top 10
display(df_all_country_28day_incident_rate.orderBy('Incident_Rate', ascending=False).take(10))
```

▸ (5) Spark Jobs

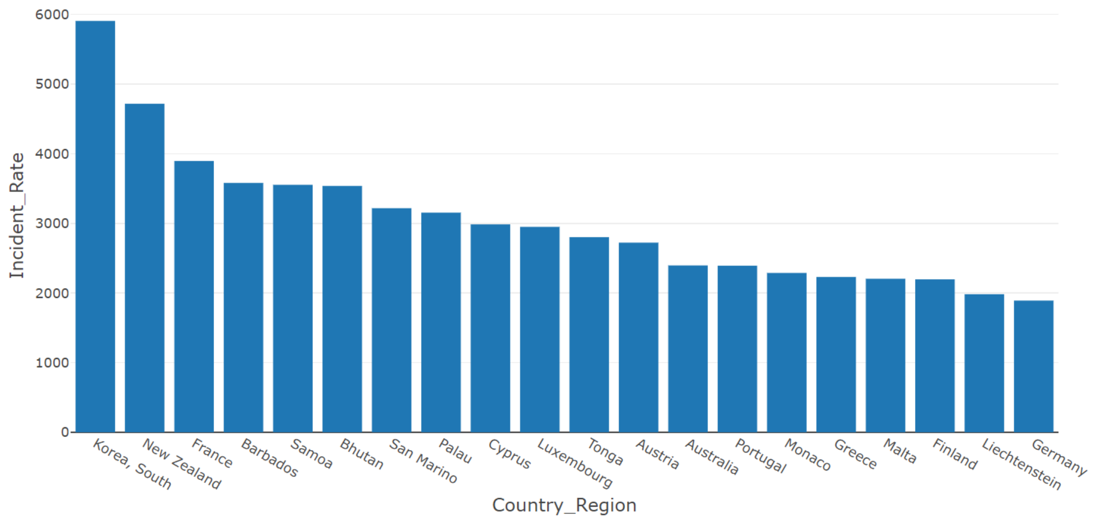
	Country_Region	Confirmed_28day	Population	Incident_Rate
1	Korea, South	3028332	51269183	5906.729584514736
2	New Zealand	228322	4839692	4717.696911291049
3	France	2653897	68128061	3895.4535929035173
4	Barbados	10289	287371	3580.3891137240707
5	Samoa	6970	196130	3553.765359710396
6	Bhutan	27292	771612	3537.0108292768905
7	San Marino	1092	33938	3217.632152749131
8	Palau	568	18008	3154.153709462461
9	Cyprus	36046	1207361	2985.5196581635482
10	Luxembourg	18466	625976	2949.9533528441984

Showing all 10 rows.

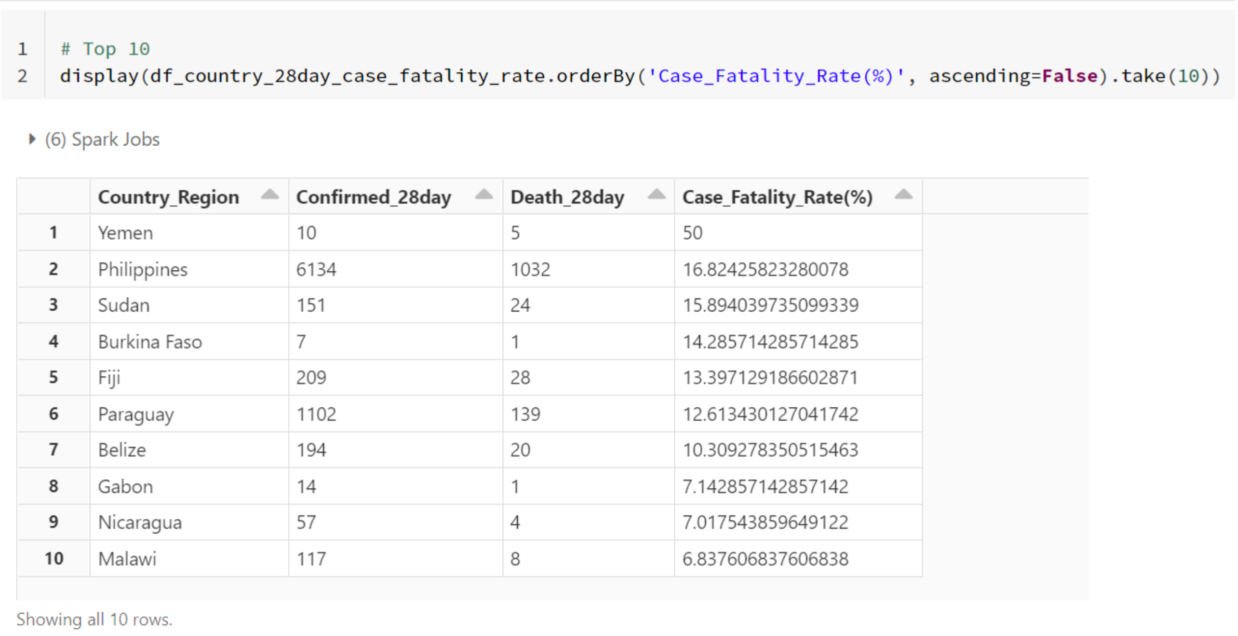


```
display(df_all_country_28day_incident_rate.orderBy('Incident_Rate', ascending=False).take(20))
```

▸ (5) Spark Jobs



Top Countries with Highest 28-Day COVID-19 Case Fatality Rate

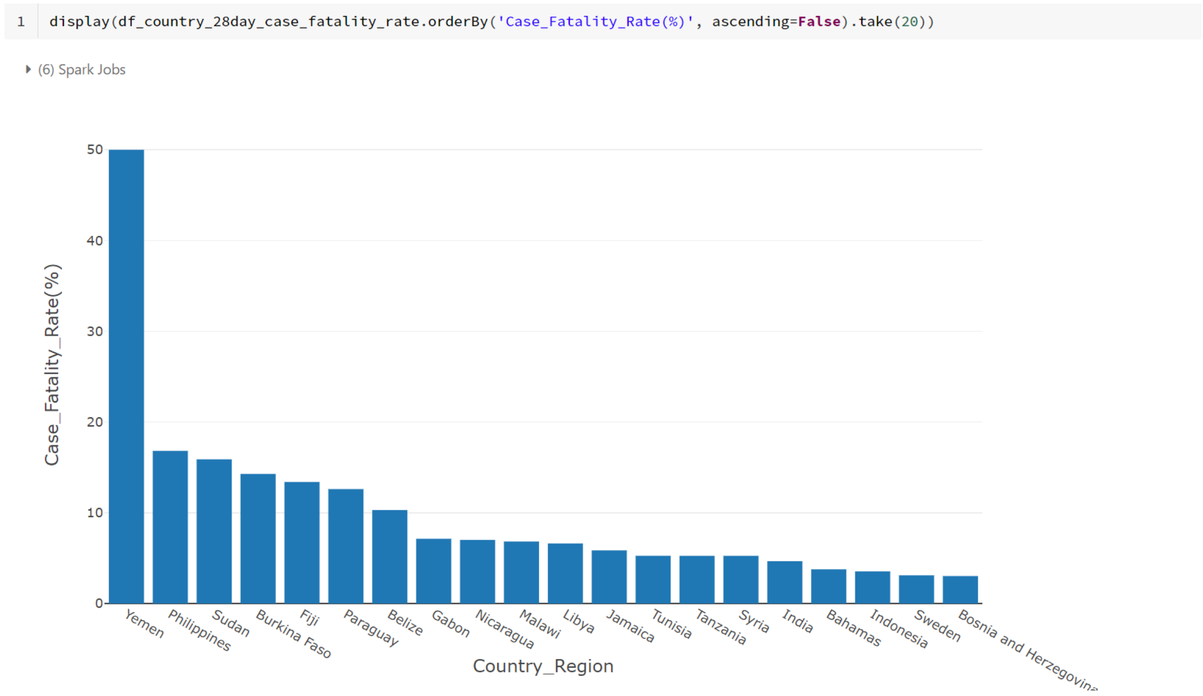


```
# Top 10
display(df_country_28day_case_fatality_rate.orderBy('Case_Fatality_Rate(%)', ascending=False).take(10))
```

▸ (6) Spark Jobs

	Country_Region	Confirmed_28day	Death_28day	Case_Fatality_Rate(%)
1	Yemen	10	5	50
2	Philippines	6134	1032	16.82425823280078
3	Sudan	151	24	15.894039735099339
4	Burkina Faso	7	1	14.285714285714285
5	Fiji	209	28	13.397129186602871
6	Paraguay	1102	139	12.613430127041742
7	Belize	194	20	10.309278350515463
8	Gabon	14	1	7.142857142857142
9	Nicaragua	57	4	7.017543859649122
10	Malawi	117	8	6.837606837606838

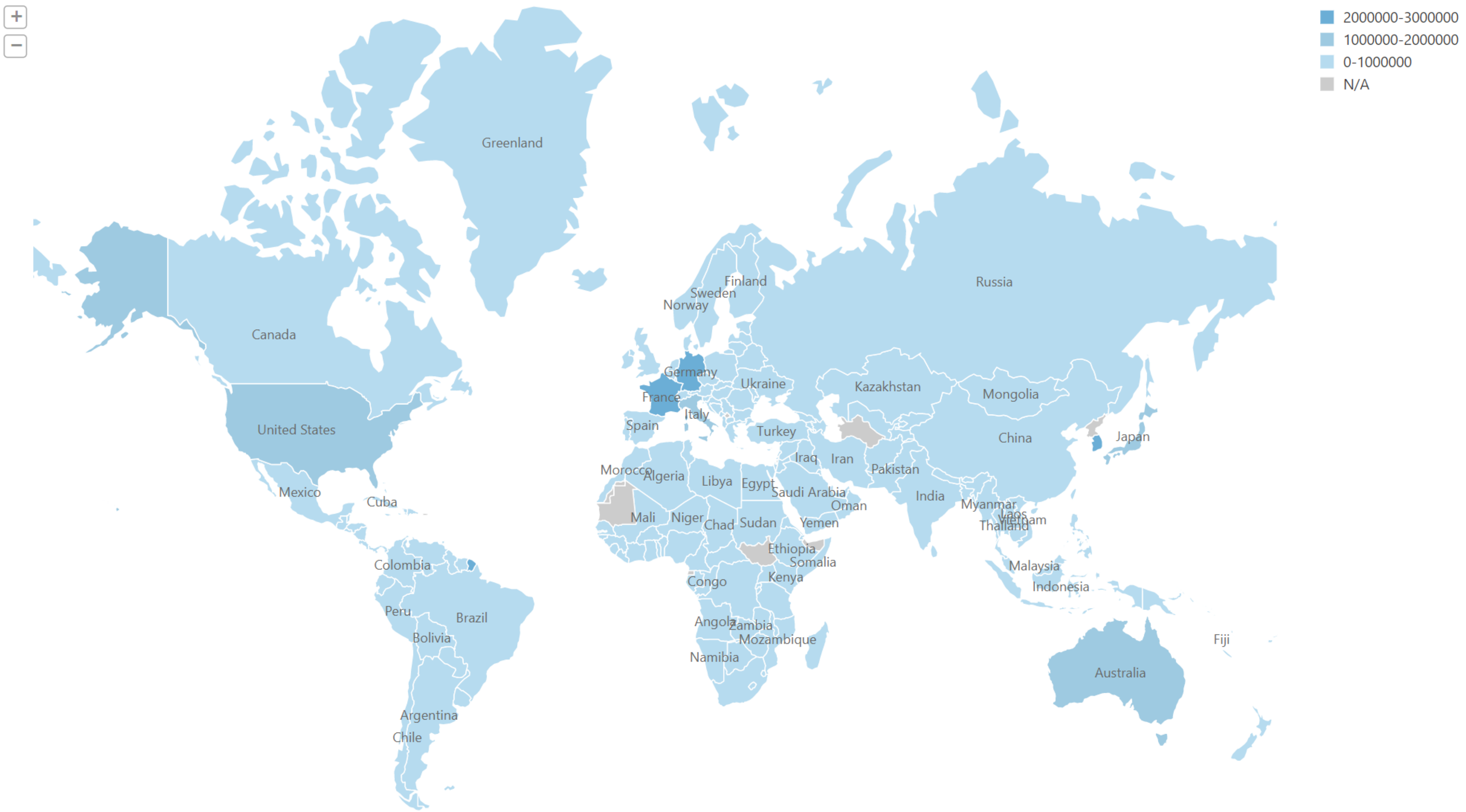
Showing all 10 rows.



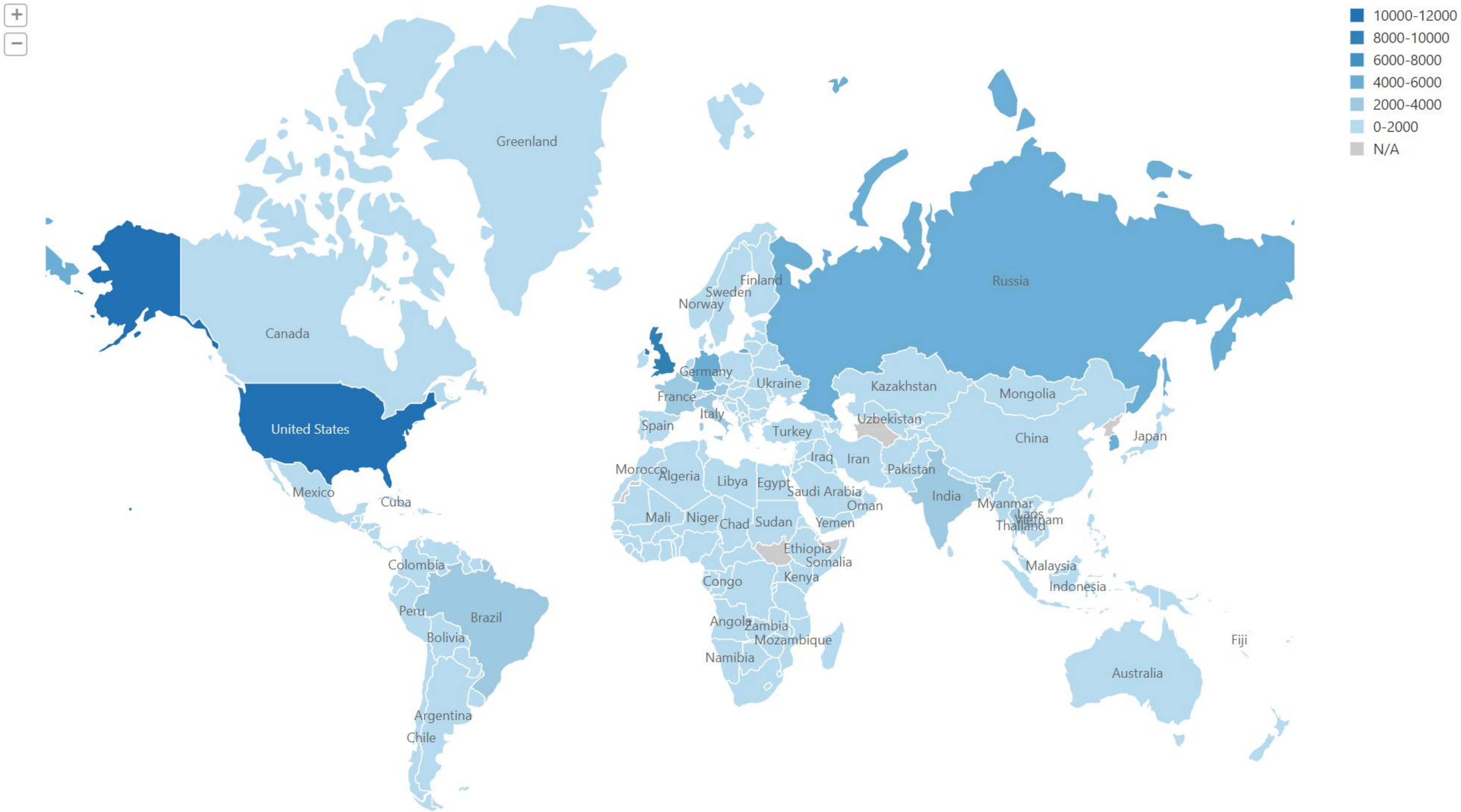
```
display(df_country_28day_case_fatality_rate.orderBy('Case_Fatality_Rate(%)', ascending=False).take(20))
```

▸ (6) Spark Jobs

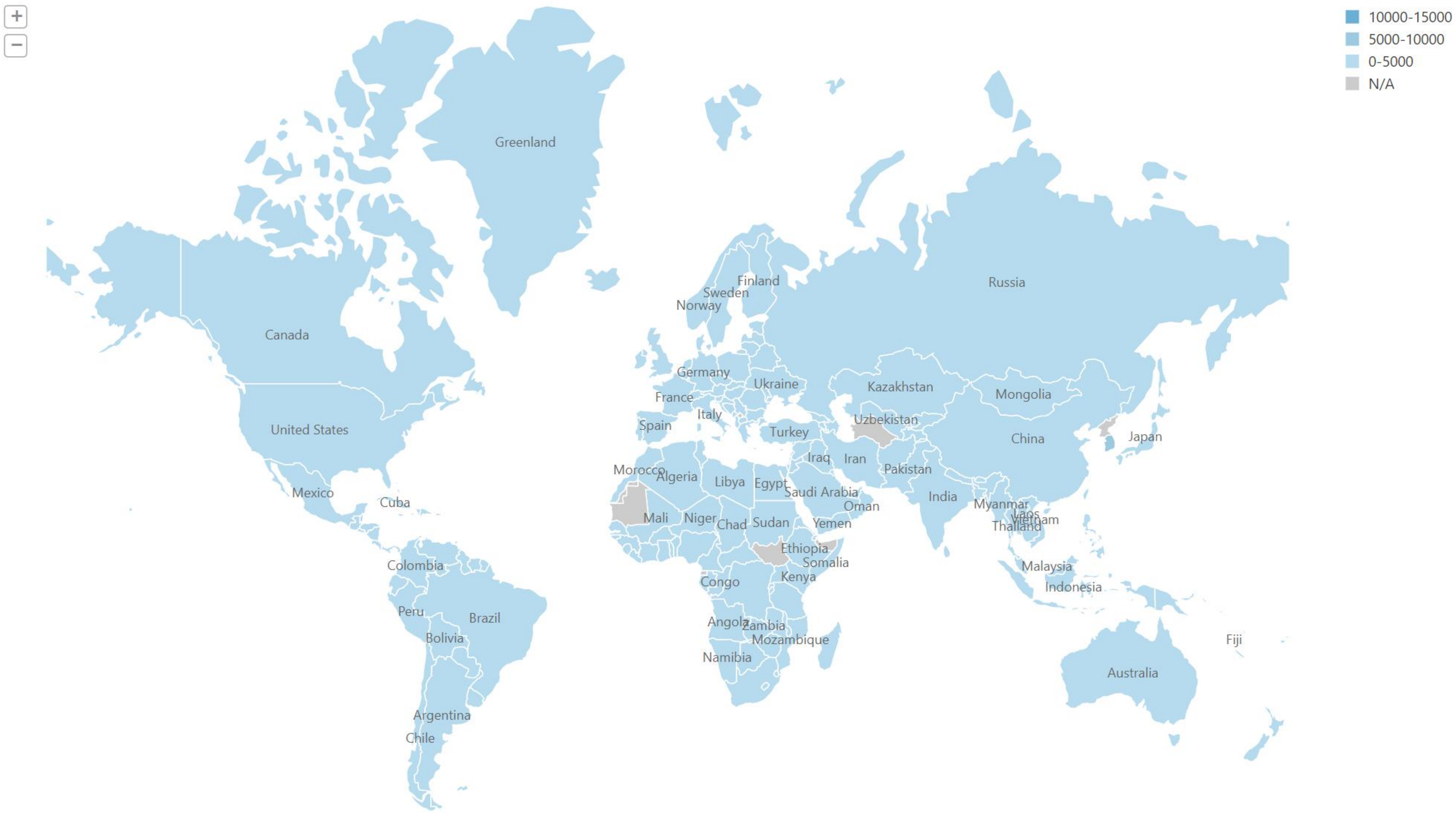
Global Distribution of 28-Day Confirmed COVID-19 Cases



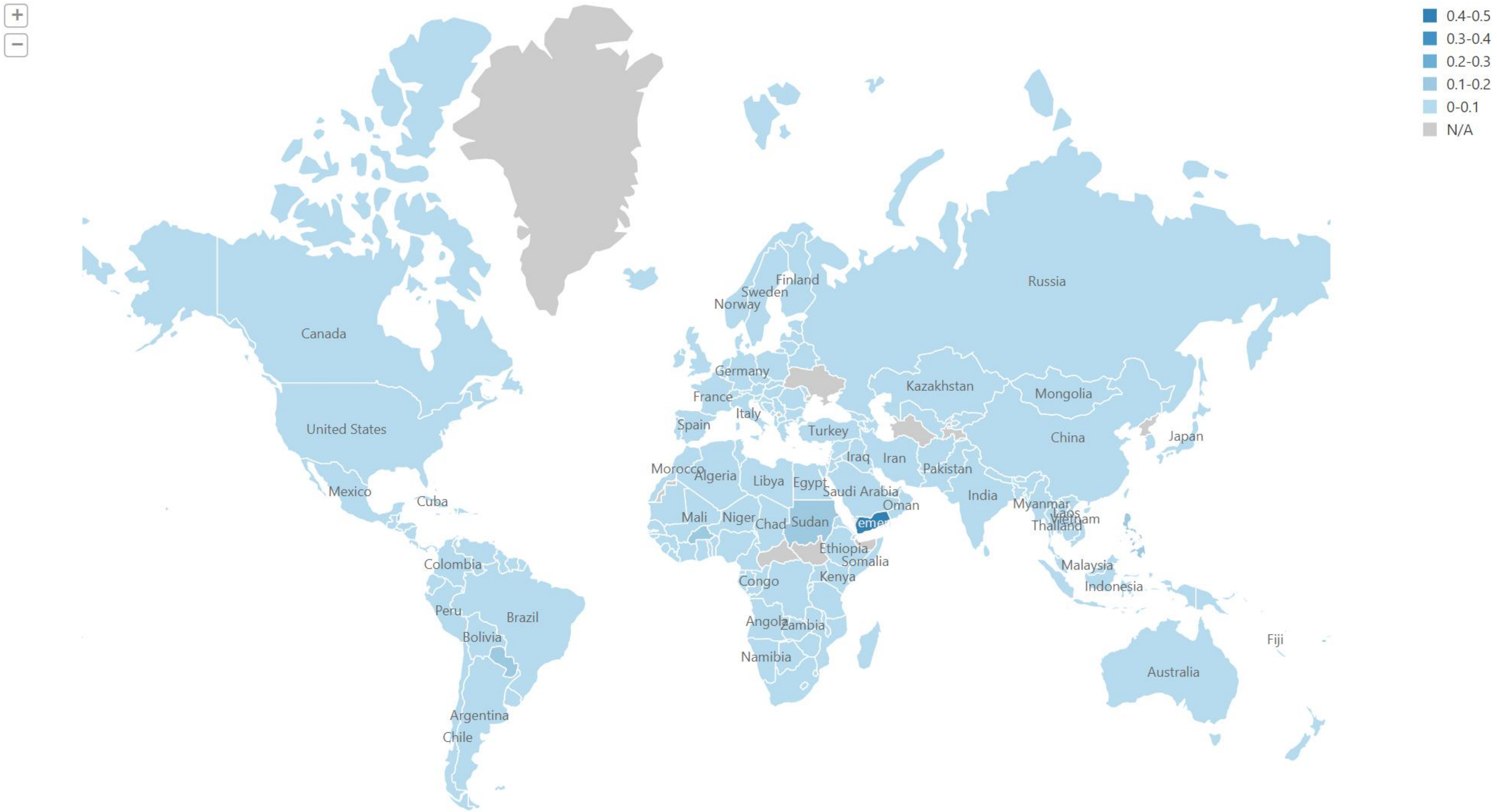
Global Distribution of 28-Day COVID-19 Death Cases



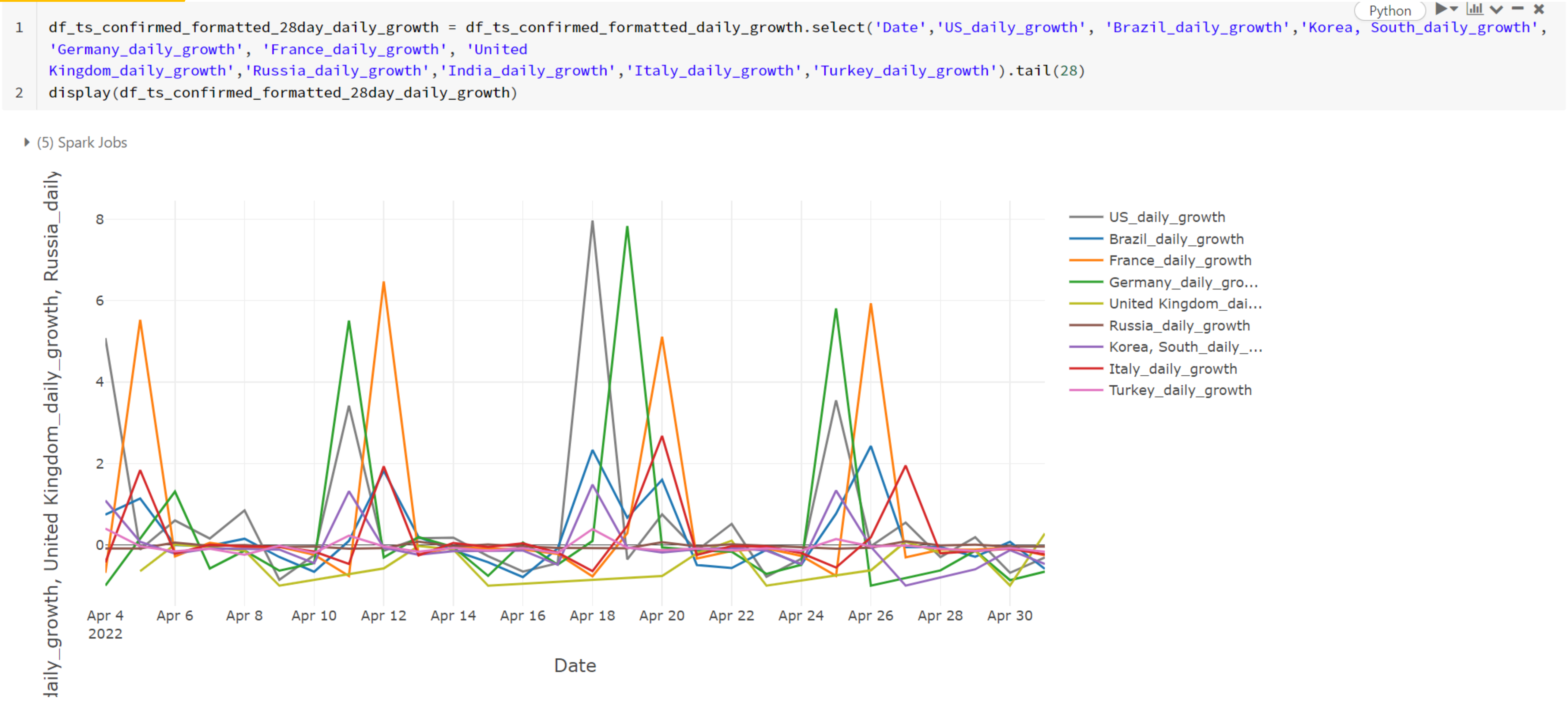
Global Distribution of 28-Day COVID-19 Incident Rate



Global Distribution of 28-Day COVID-19 Case Fatality Rate



Appendix



```python
df_ts_confirmed_formatted_28day_daily_growth = df_ts_confirmed_formatted_daily_growth.select('Date','US_daily_growth', 'Brazil_daily_growth','Korea, South_daily_growth', 'Germany_daily_growth', 'France_daily_growth', 'United Kingdom_daily_growth','Russia_daily_growth','India_daily_growth','Italy_daily_growth','Turkey_daily_growth').tail(28)
display(df_ts_confirmed_formatted_28day_daily_growth)
```

(5) Spark Jobs