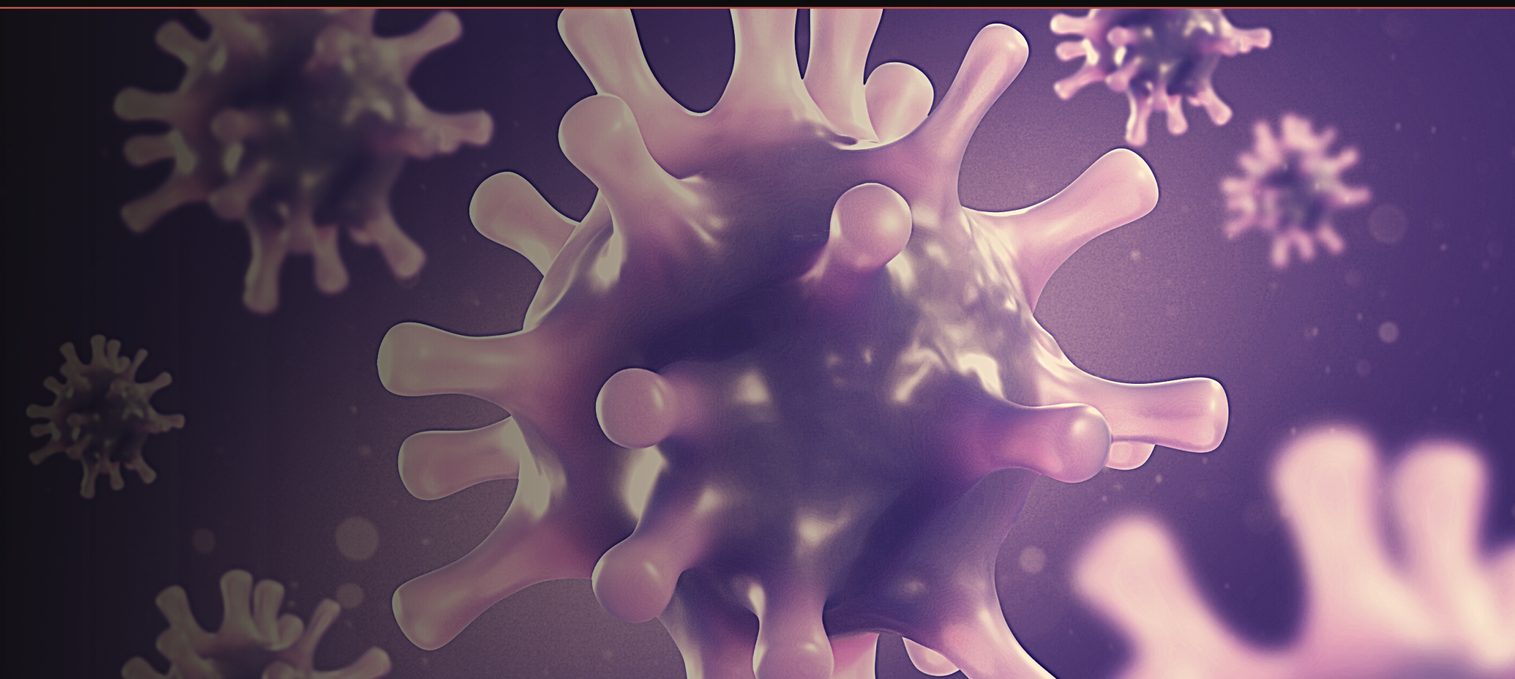
Covid-19 Analysis and Forecasting
Reeti Bhagat

Data Science intensive capstone project May26th,2020 Cohort



Thanks to Springboard mentors

Ash Yousefi

Springboard

The Problem

- Covid -19 is highly contagious disease that has spread worldwide leading to global health crises.

- It is highly contagious disease and the exact cause is not known. It is global pandemic (WHO) .

- It has affected more than 20 million and killed 0.9 million people in world.

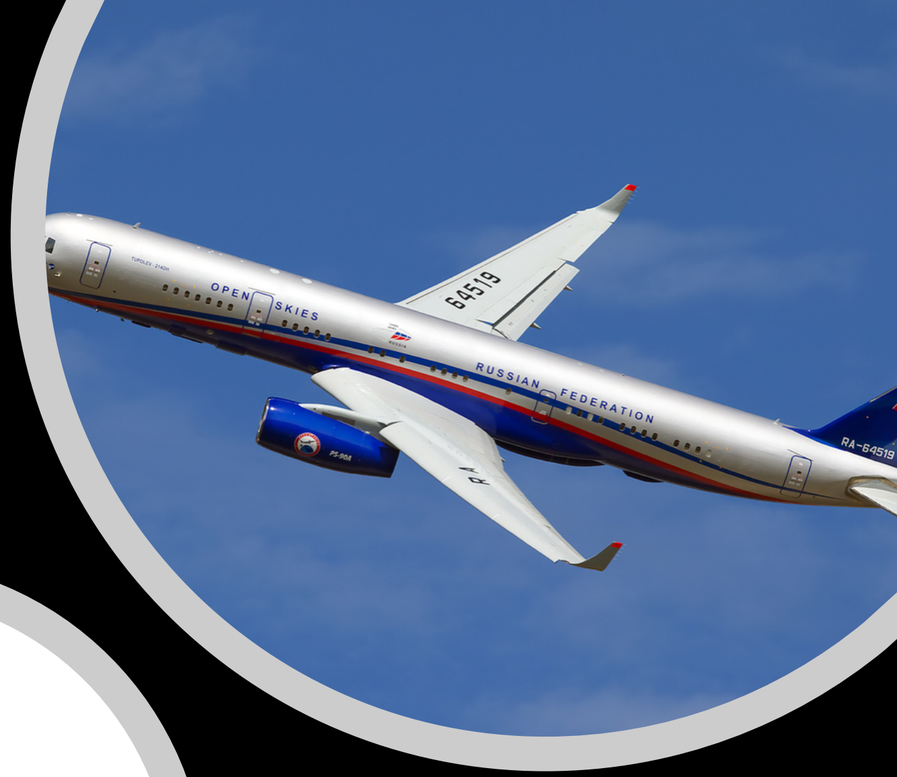
Why should be Concerned?

How does Covid-19 affect different countries?

How can we find the projections regarding the number of cases?

Who might Cares??

- Airlines
- Travel Agencies and Hotels
- Entertainment
- Employment services
- Education



Data Information

- **2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository by Johns Hopkins CSSE (LINK)**

- Dataset consists of time-series data from 22 JAN 2020 to AUG 8,2020.

- **Time-series dataset:**
- time_series_covid19_confirmed_global.csv (Link Raw File)
- time_series_covid19_deaths_global (Link Raw File)

worldometers.info/coronavirus/

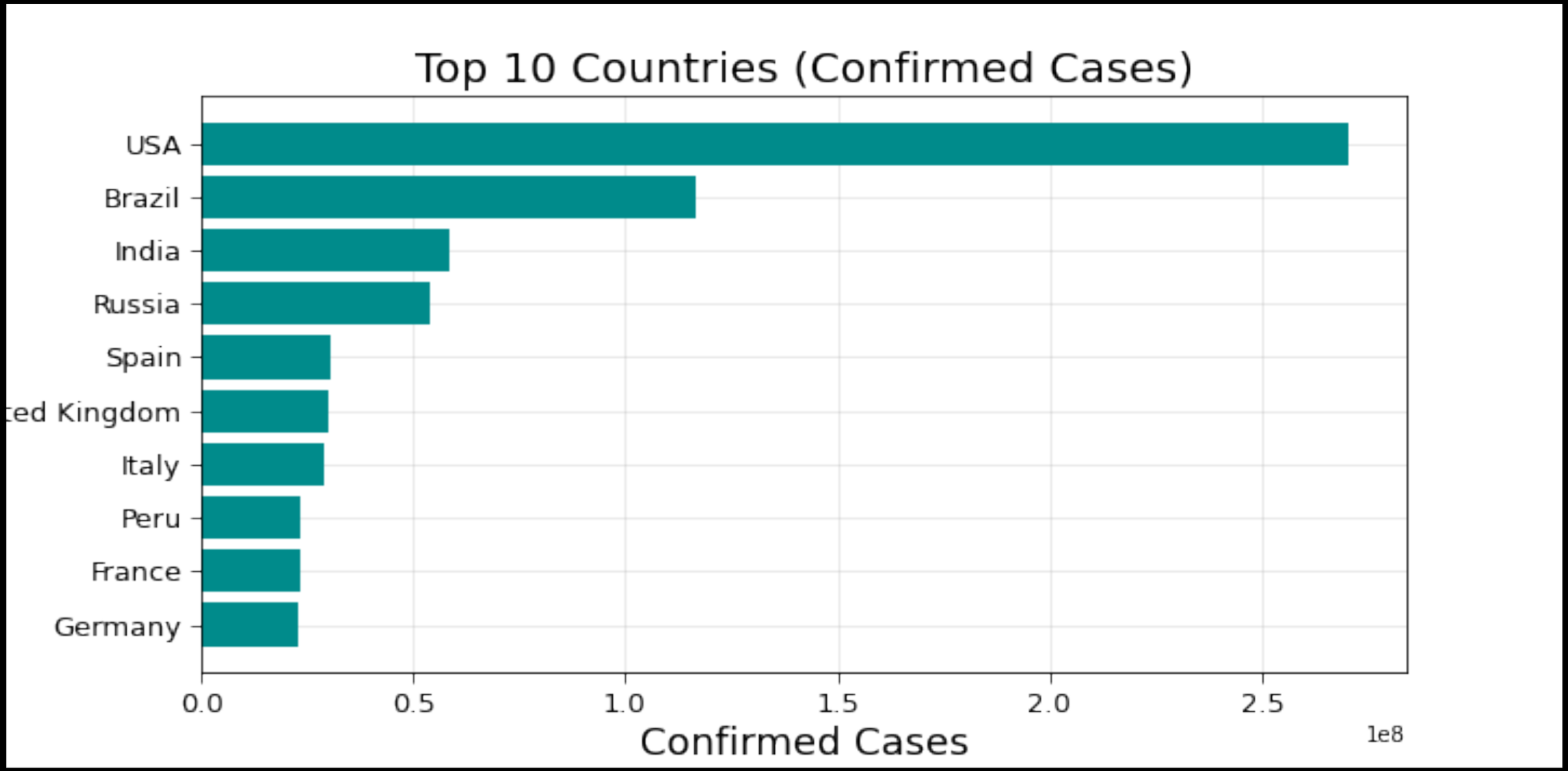
Data Exploration

https://github.com/reetibhagat/capstone-1-covid-19/blob/master/notebooks/Exploratory_data_analyis_capstone_1.ipynb

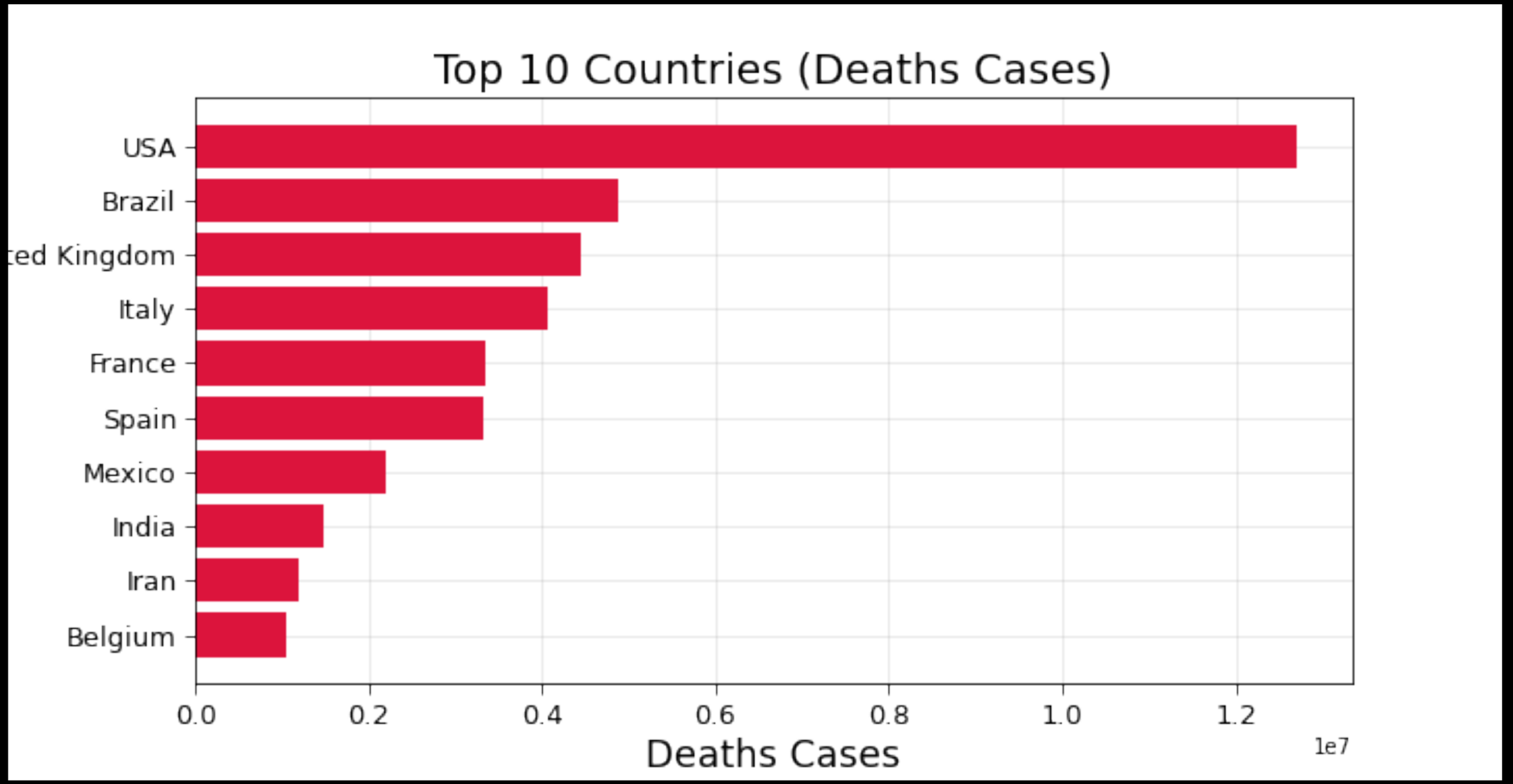
General Analysis of Data(Continent wise Data)

continent	Confirmed	Deaths	Recovered	Mortality Rate	Recovery Rate	Active Cases
Africa	37709406	904389	19805801	23529.65	283044.01	16999216
Asia	219219516	5601618	144850378	15027.05	336605.64	68767520
Australia	1337210	16006	982688	419.58	38472.80	338516
Europe	249950948	20723423	120922012	30722.22	313984.74	108305513
North America	312311311	16014189	90841316	13642.47	148984.47	205455806
Others	889051	13722	444799	4978.22	59165.06	430530
South America	186568641	7142013	116023973	6994.25	66492.83	63402655

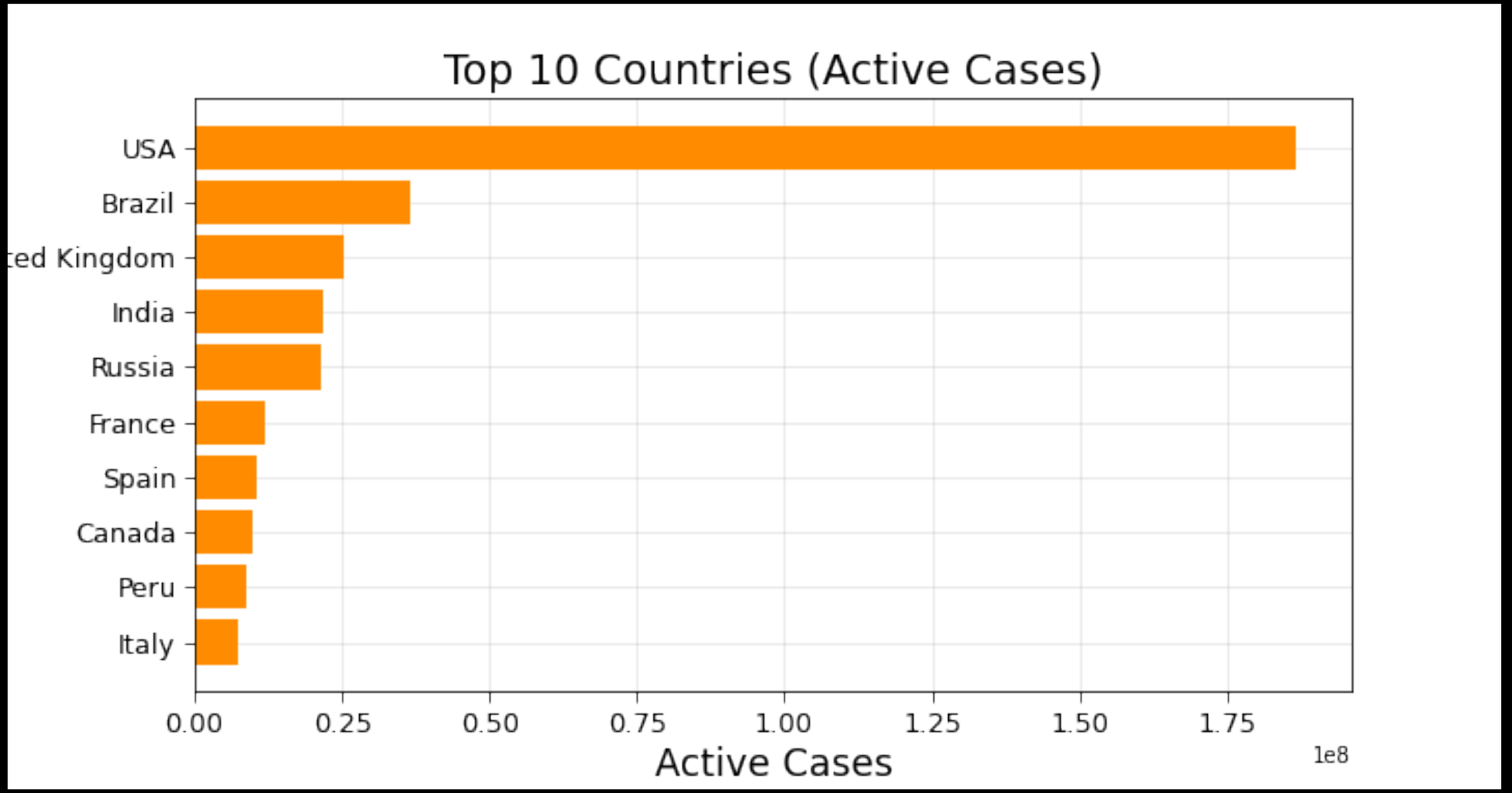
Top 10 COUNTRIES(CONFIRMED CASES)



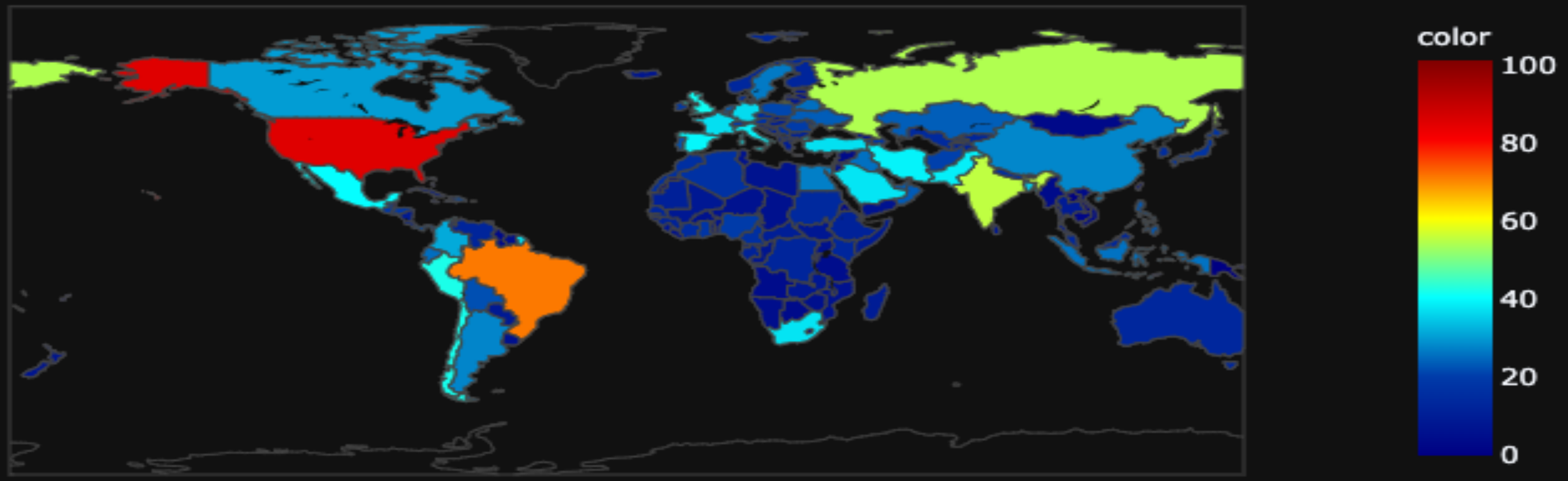
Top 10 Countries(Deaths Cases)



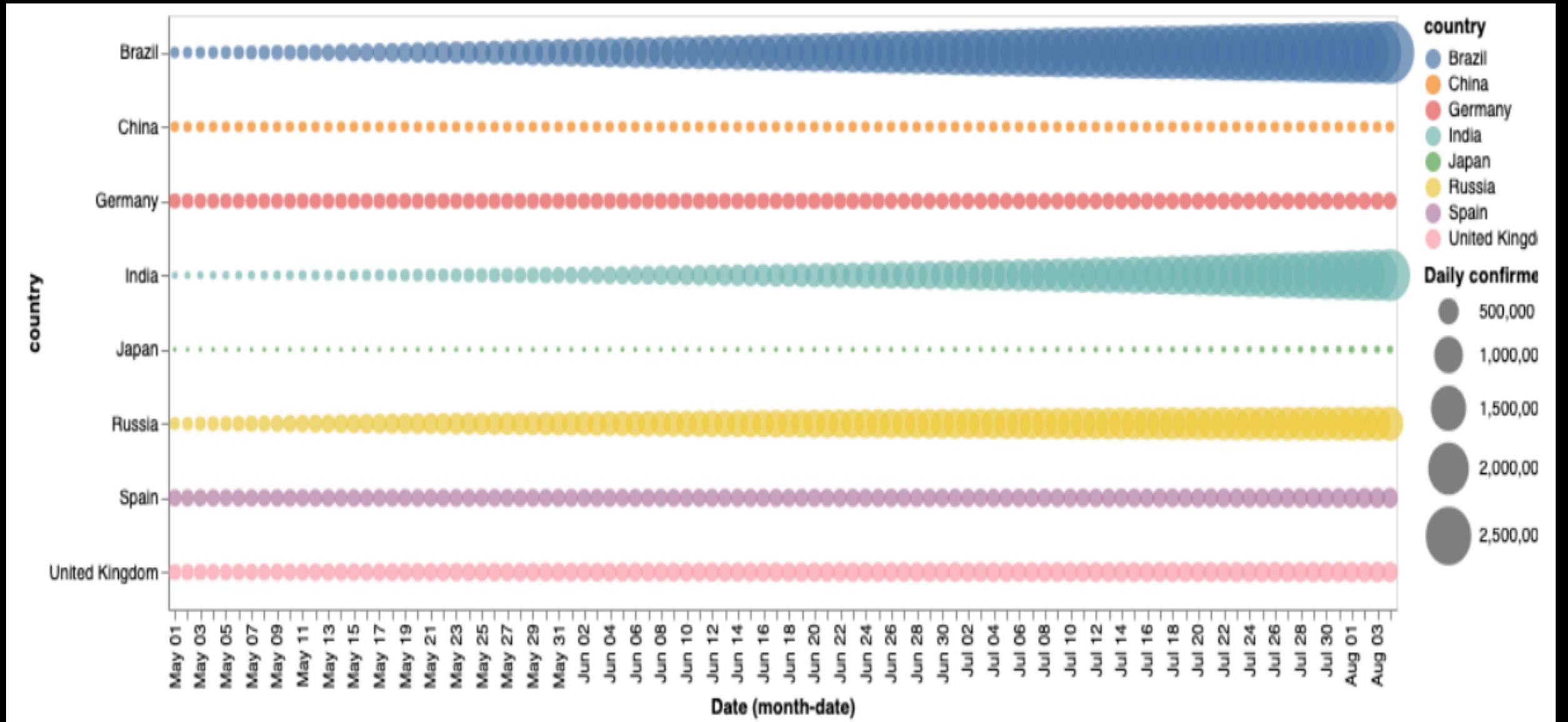
Top 10 Countries(Active Cases)

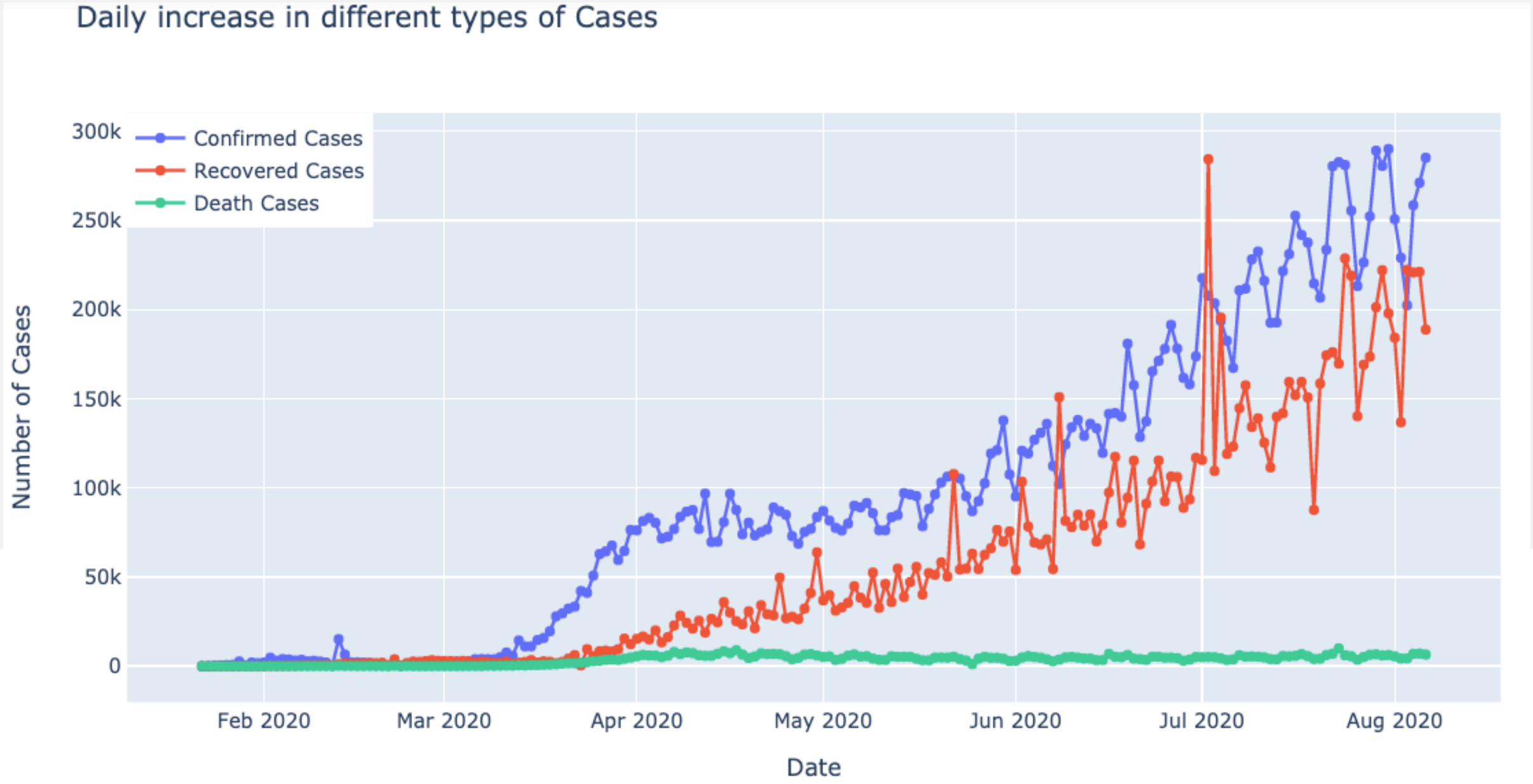


Covid-19:Progression of Spread

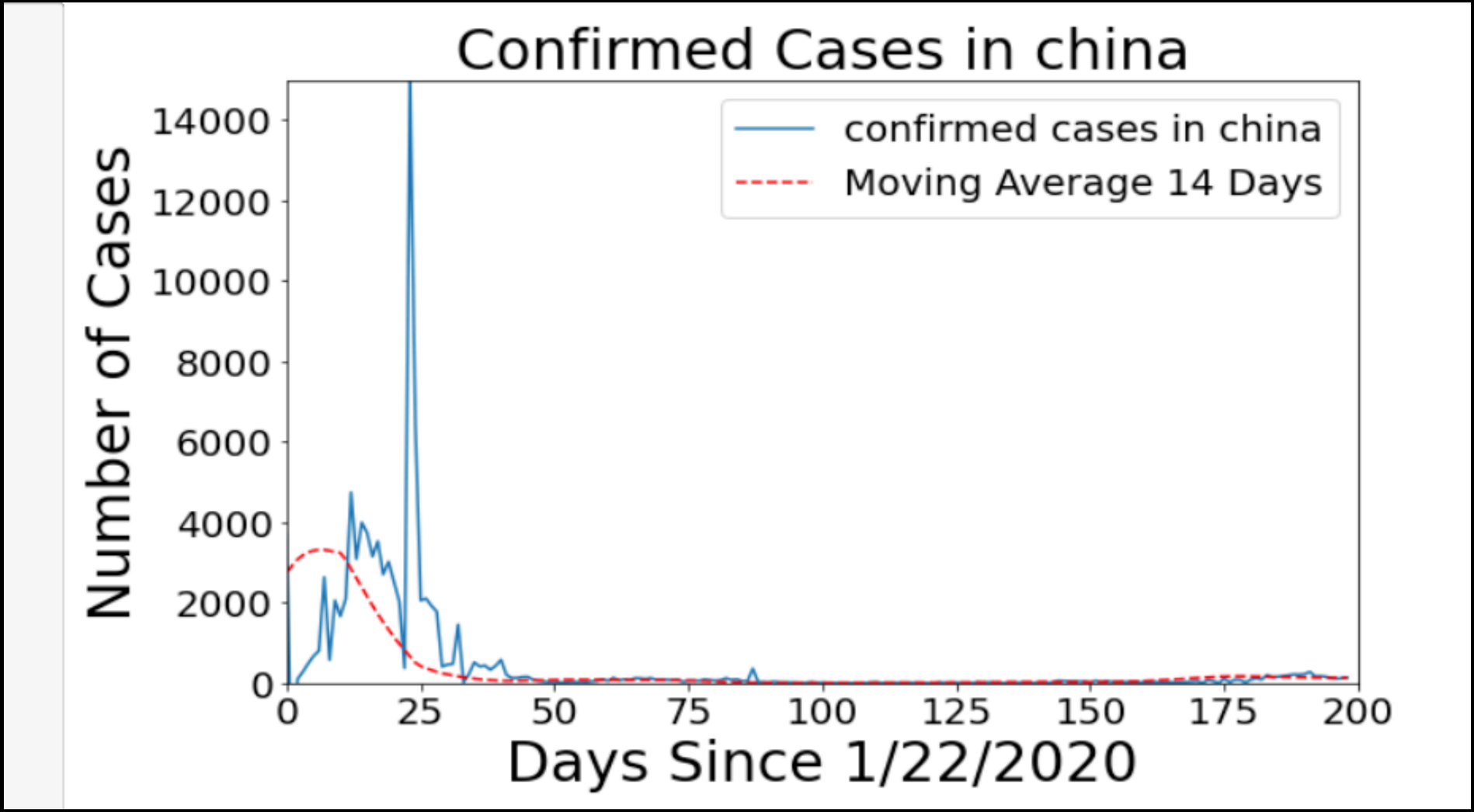


Comparison of spread of Covid-19 in different countries

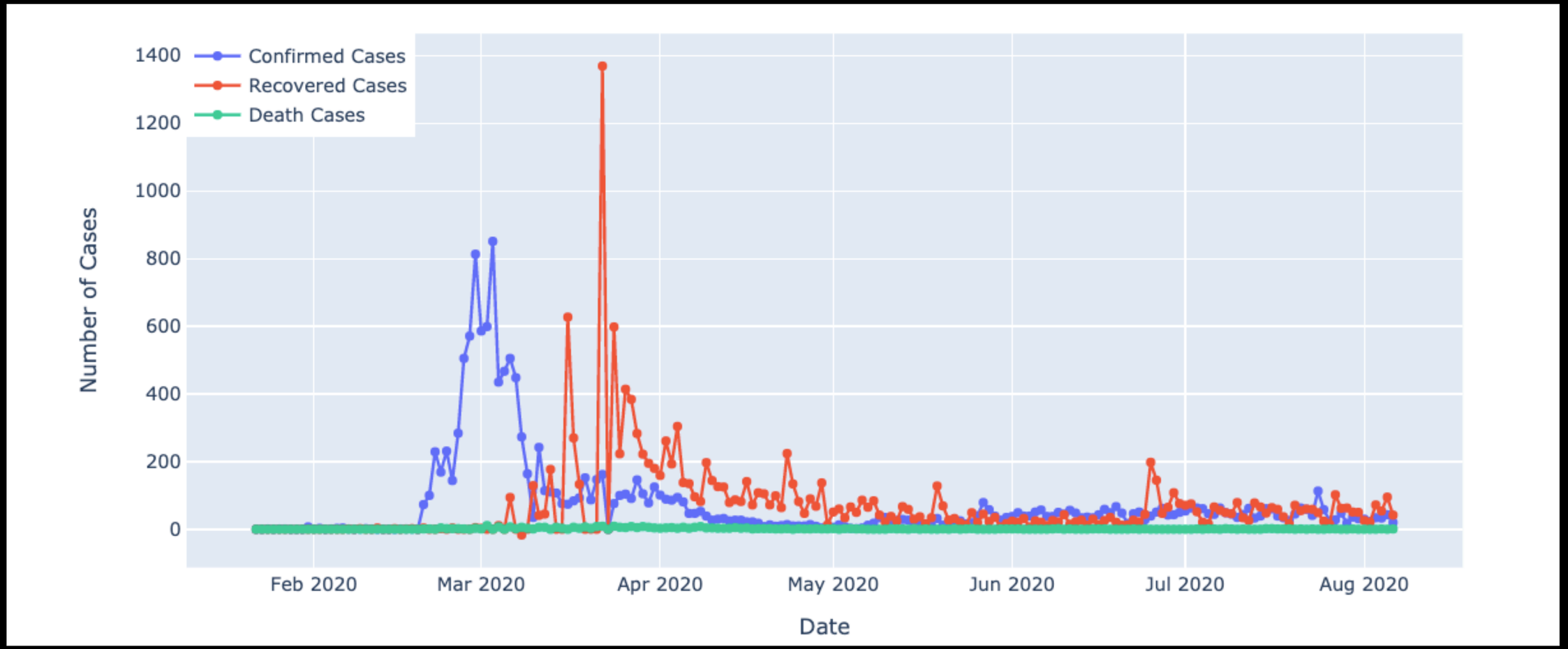



Daily increase in different types of Cases
Confirmed Cases
Recovered Cases
Death Cases
300k
250k
200k
150k
100k
50k
0
Number of Cases
Feb 2020
Mar 2020
Apr 2020
May 2020
Jun 2020
Jul 2020
Aug 2020
Date

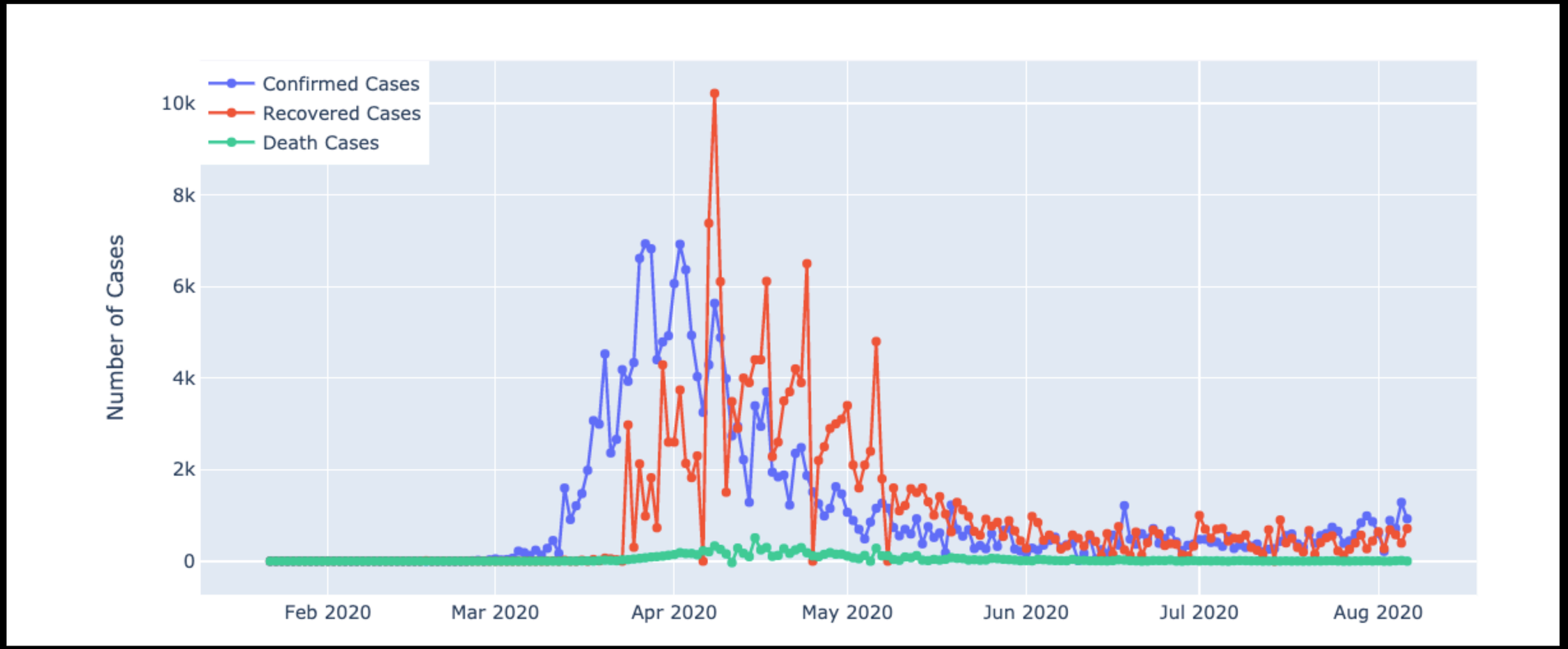
Confirmed Cases in China



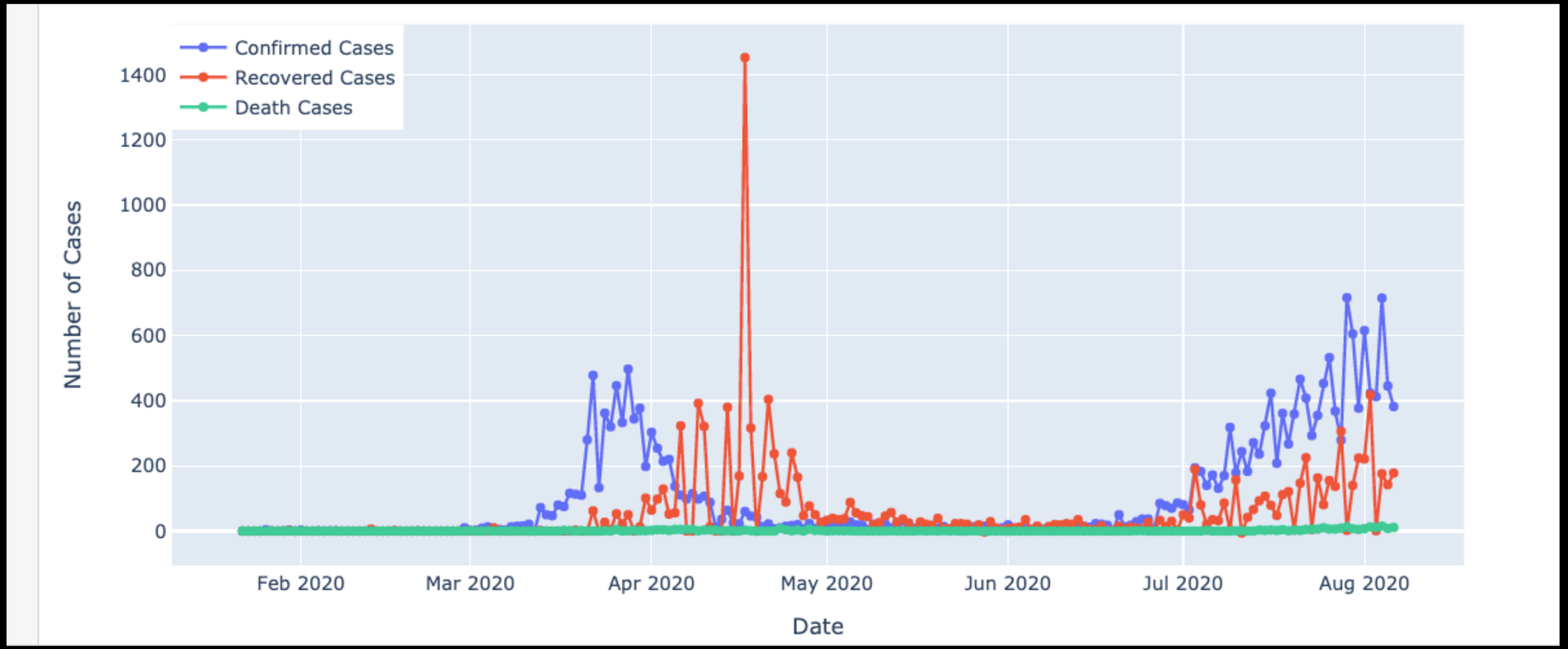
South Korea


Confirmed Cases
Recovered Cases
Death Cases
1400
1200
1000
800
600
400
200
0
Number of Cases
Feb 2020
Mar 2020
Apr 2020
May 2020
Jun 2020
Jul 2020
Aug 2020
Date

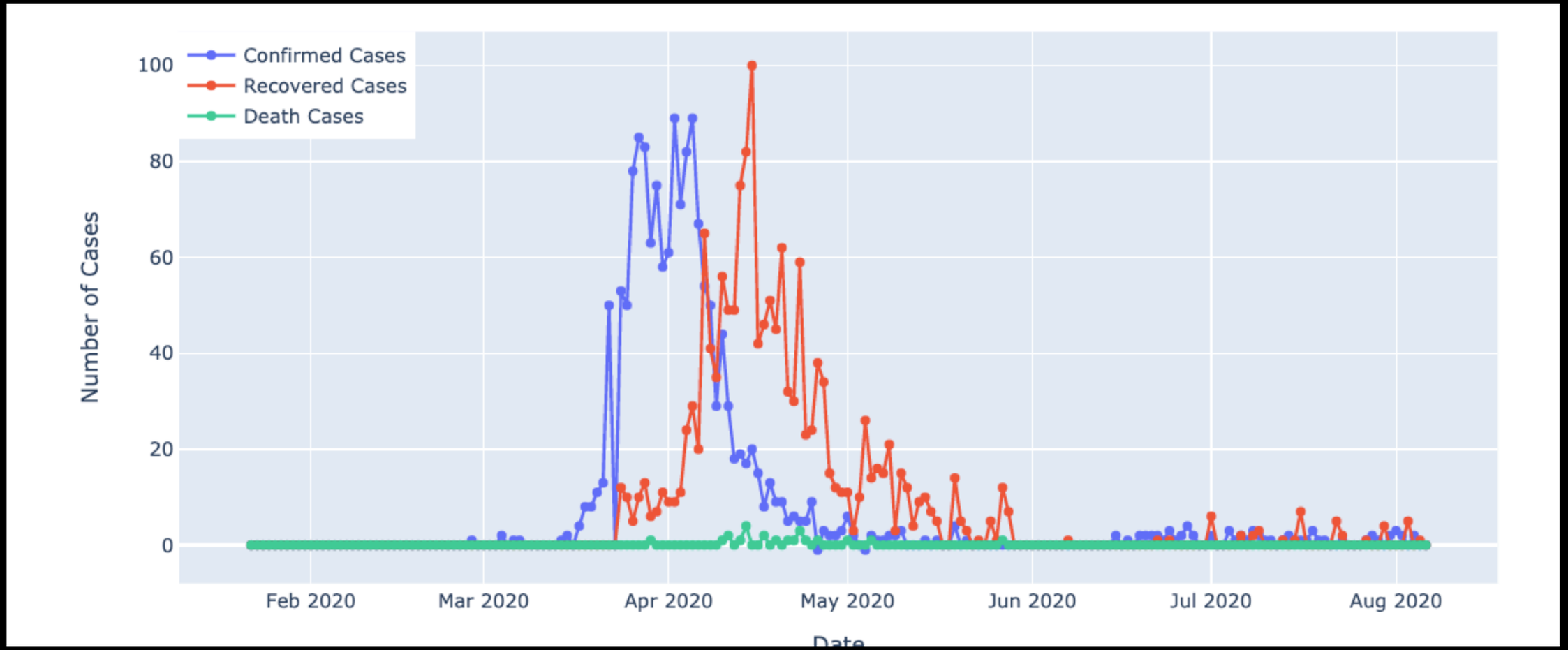
Germany

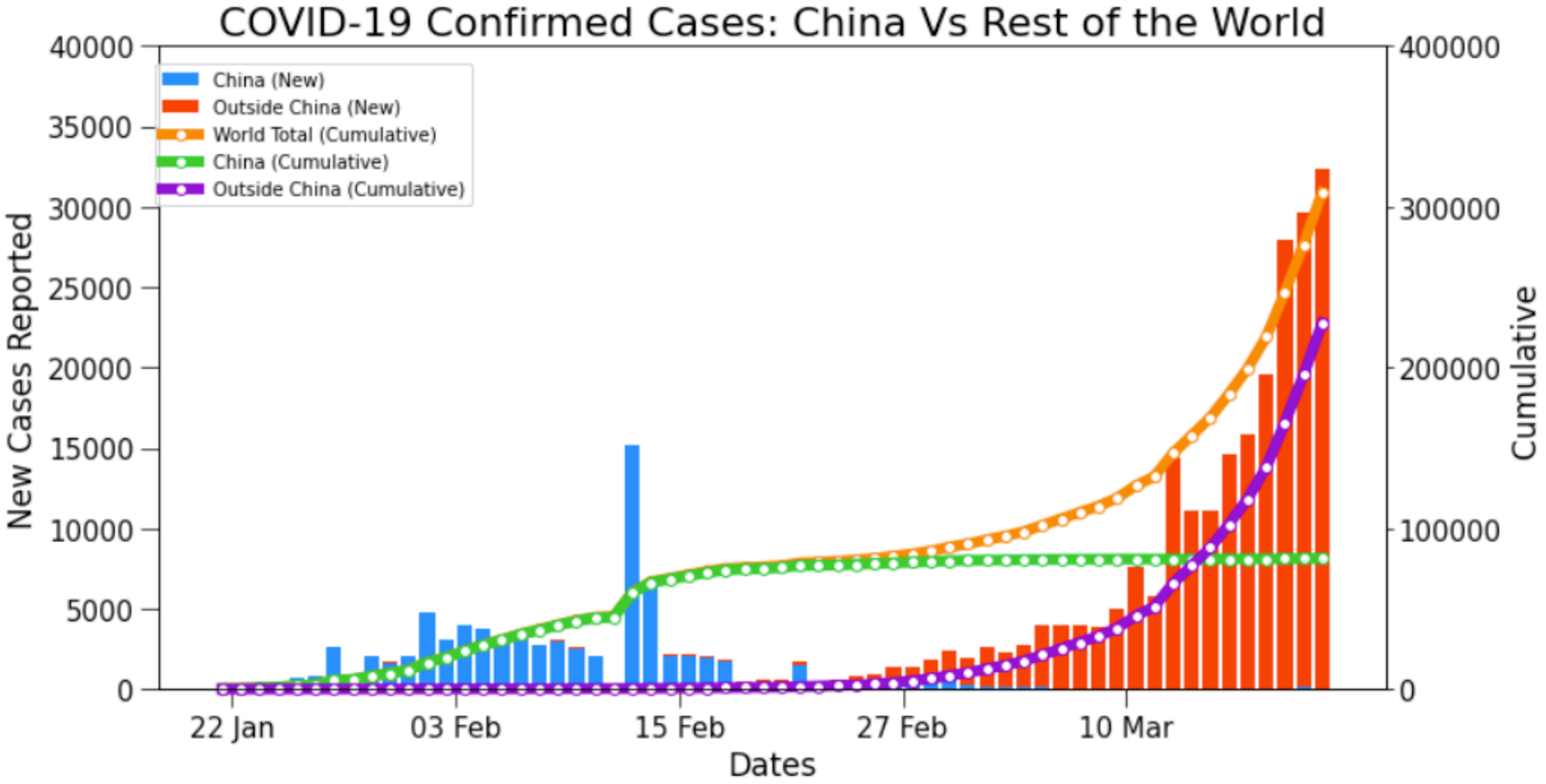

Confirmed Cases
Recovered Cases
Death Cases
Number of Cases
10k
8k
6k
4k
2k
0
Feb 2020
Mar 2020
Apr 2020
May 2020
Jun 2020
Jul 2020
Aug 2020

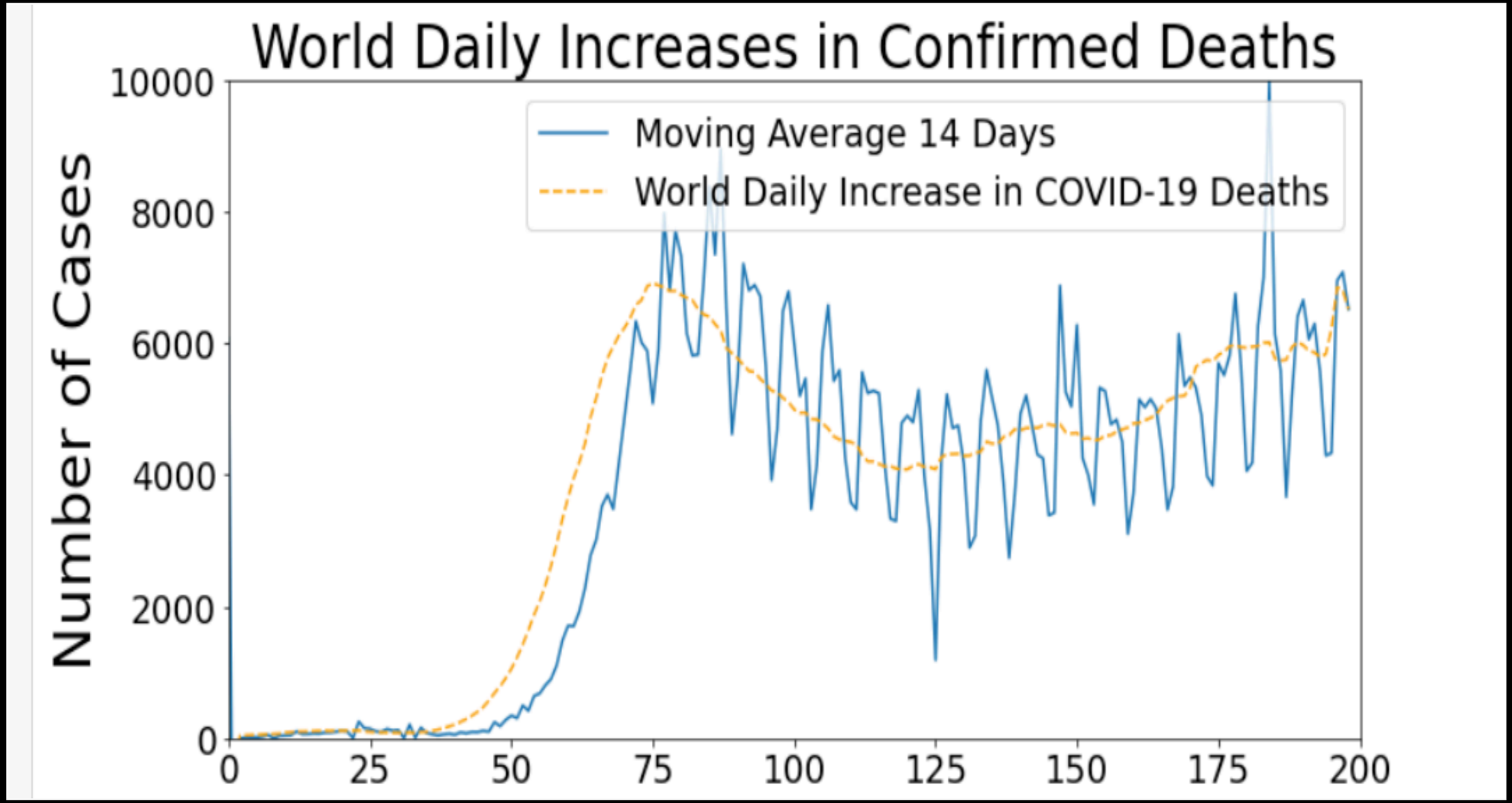
Australia


Confirmed Cases
Recovered Cases
Death Cases
Number of Cases
0
200
400
600
800
1000
1200
1400
Feb 2020
Mar 2020
Apr 2020
May 2020
Jun 2020
Jul 2020
Aug 2020
Date

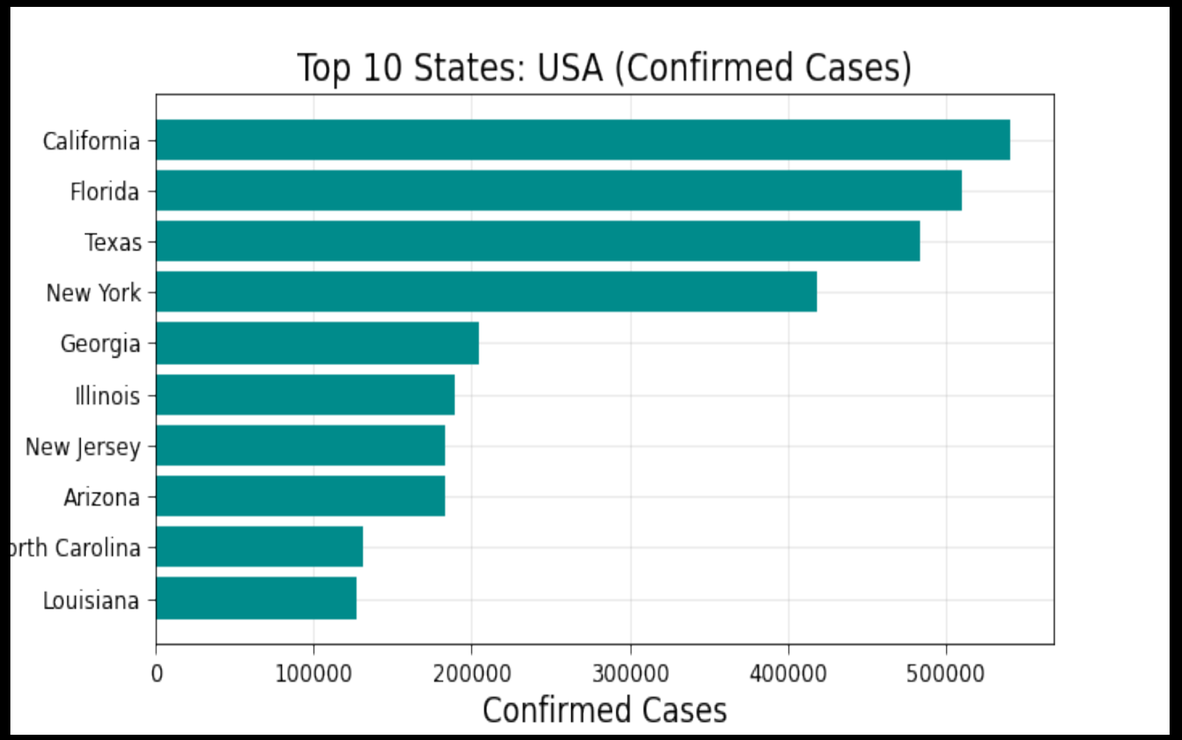
New Zealand

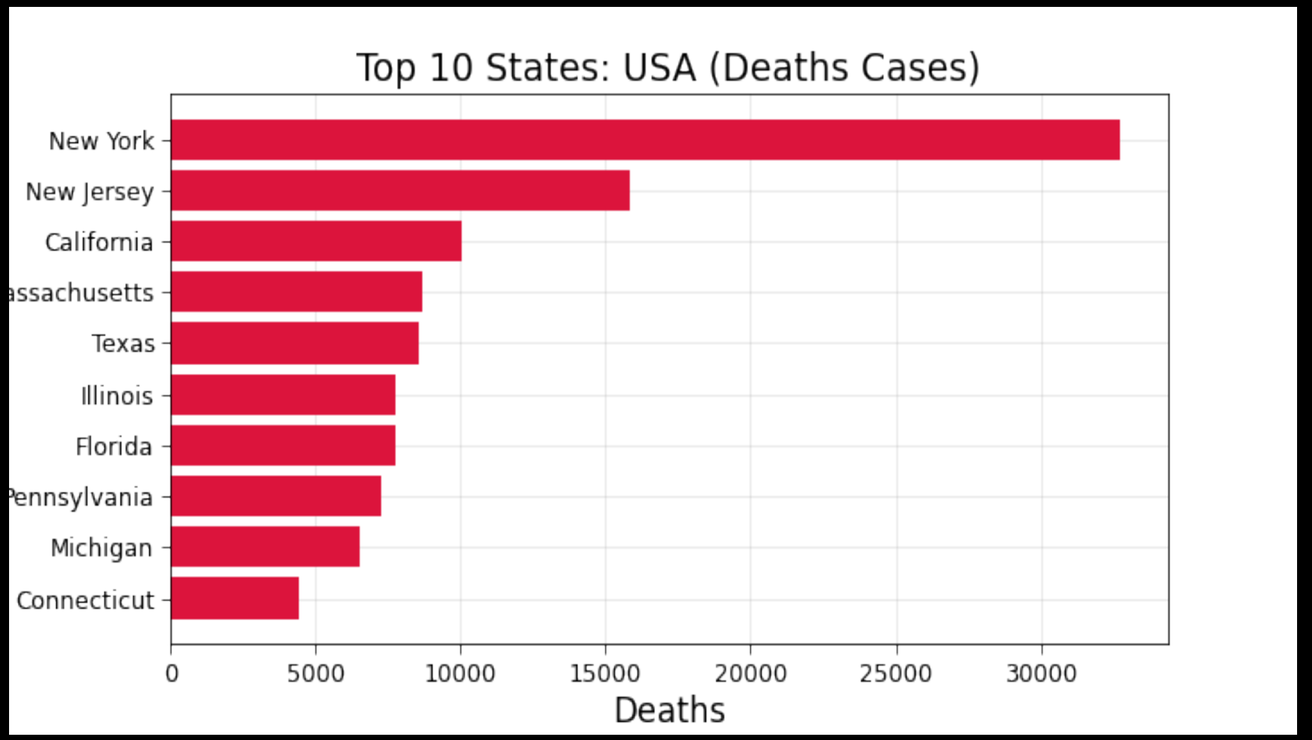

Confirmed Cases
Recovered Cases
Death Cases
Number of Cases
100
80
60
40
20
0
Feb 2020
Mar 2020
Apr 2020
May 2020
Jun 2020
Jul 2020
Aug 2020
Date


COVID-19 Confirmed Cases: China Vs Rest of the World
China (New)
Outside China (New)
World Total (Cumulative)
China (Cumulative)
Outside China (Cumulative)
New Cases Reported
Cumulative
Dates
22 Jan
03 Feb
15 Feb
27 Feb
10 Mar
0
5000
10000
15000
20000
25000
30000
35000
40000
0
100000
200000
300000
400000


World Daily Increases in Confirmed Deaths
Moving Average 14 Days
World Daily Increase in COVID-19 Deaths
Number of Cases
10000
8000
6000
4000
2000
0
0
25
50
75
100
125
150
175
200

USA


Top 10 States: USA (Confirmed Cases)
California
Florida
Texas
New York
Georgia
Illinois
New Jersey
Arizona
orth Carolina
Louisiana
0
100000
200000
300000
400000
500000
Confirmed Cases


Top 10 States: USA (Deaths Cases)
New York
New Jersey
California
assachusetts
Texas
Illinois
Florida
Pennsylvania
Michigan
Connecticut
0
5000
10000
15000
20000
25000
30000
Deaths

Machine learning Modeling

https://github.com/reetibhagat/capstone-1-covid-19/blob/master/notebooks/Modelling.ipynb

Time Series Analysis

Modeling Overview

- **Time Series Analysis**:

  It is a series of observations taken at specified times basically at equal intervals. It is used to predict future values based on past observed values.

- **Comparison and validation of data using different Models:**

  1.EXPONENTIAL SMOOTHING

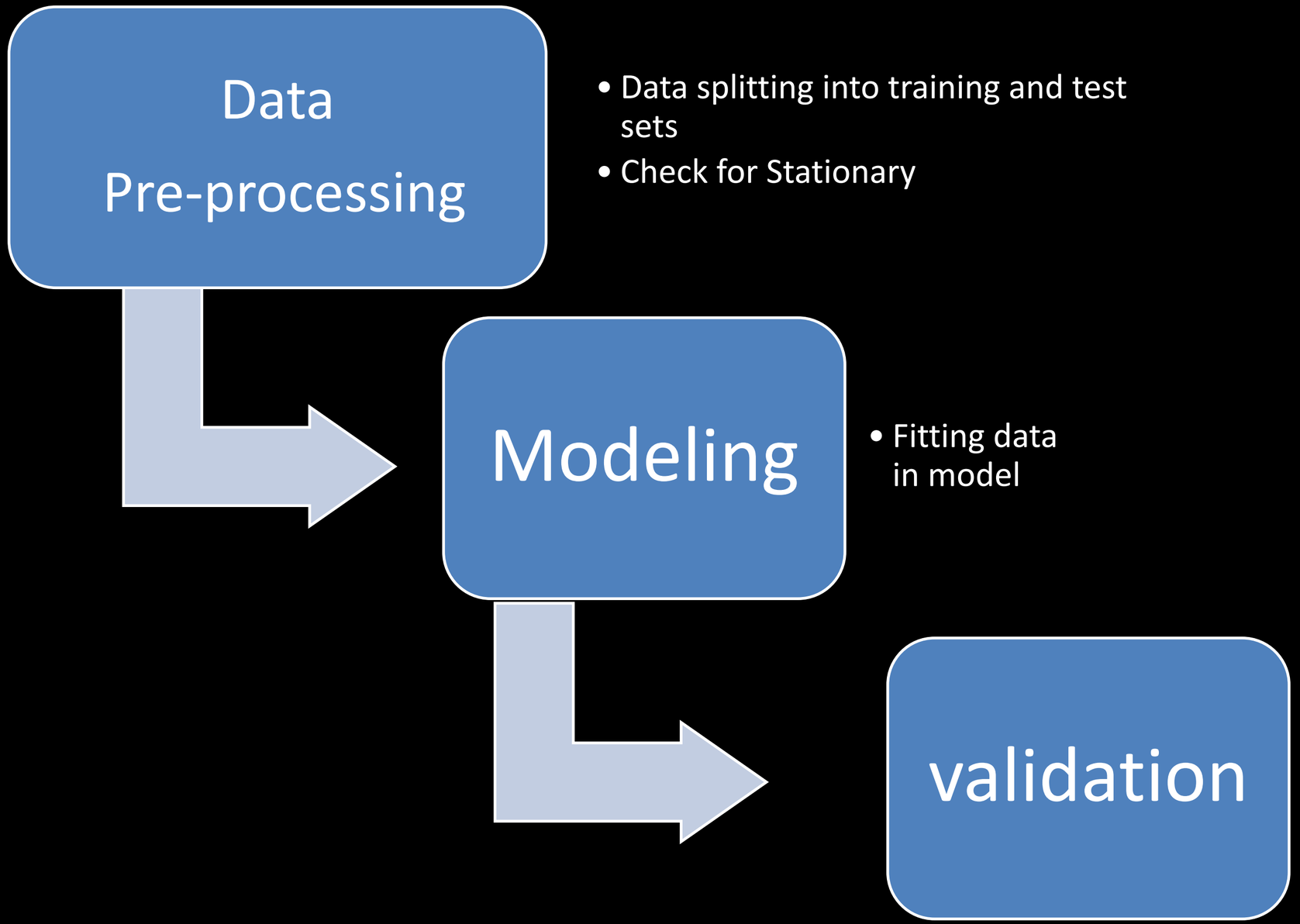
  2.ARIMA

  3.SARIMA

  4.PROPHET

- **Tools used**: Stats. Models and Scikit Learn

Modeling steps



	Model Name	Root Mean Squared Error
2	SARIMA Model	6.429346e+04
0	Holt's Winter Model	8.106599e+04
1	ARIMA Model	1.027092e+05
3	Facebook's Prophet Model	1.793965e+07

Models Comparison

- RMSE of SARIMA Model has good accuracy with RMSE 64293.45 and AIC score 3922 so I will be using SARIMA model to forecast covid19 cases.

SOME DETAILS ON BEST MODEL(SARIMA)

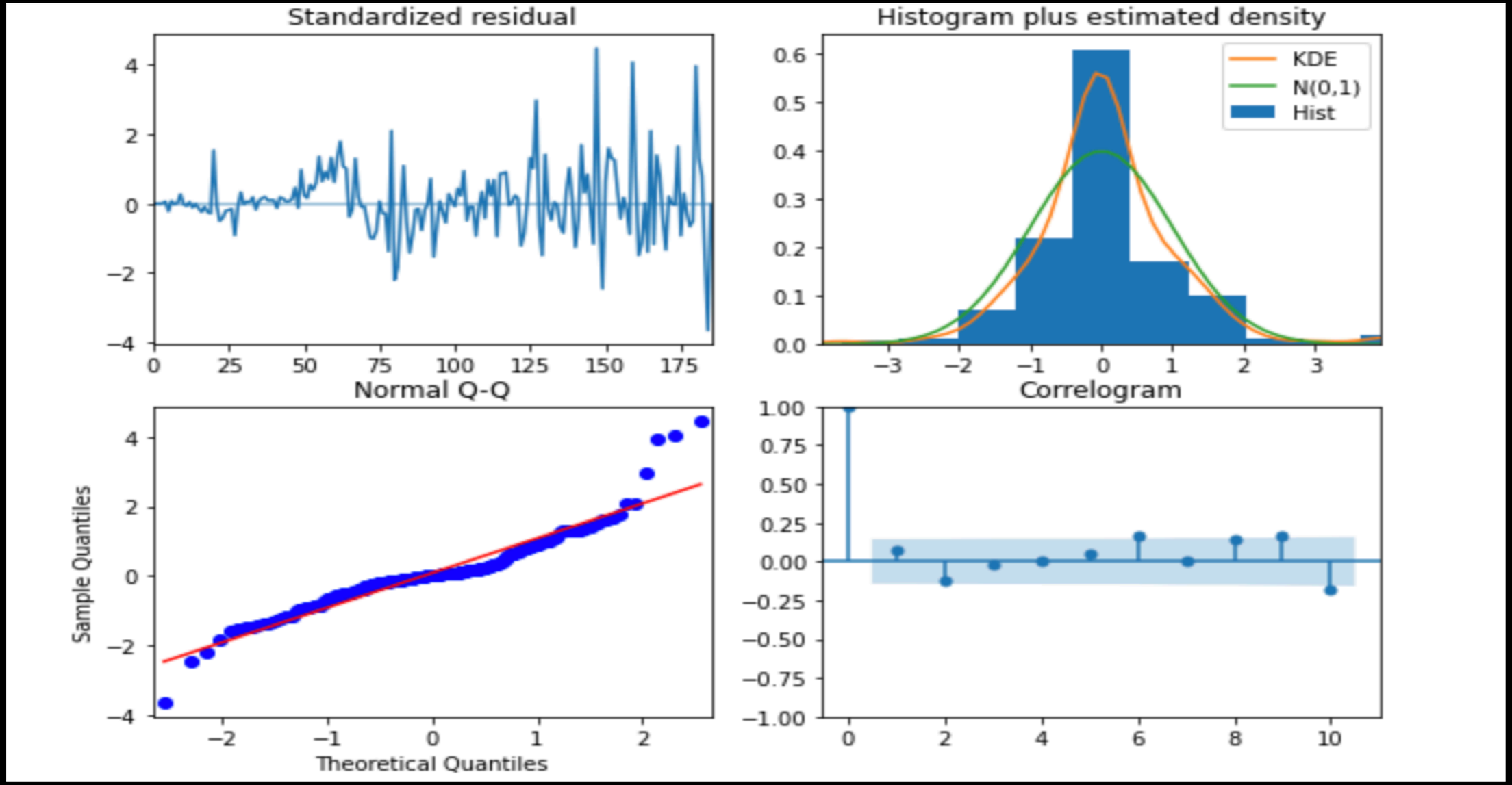
SARIMAX Results

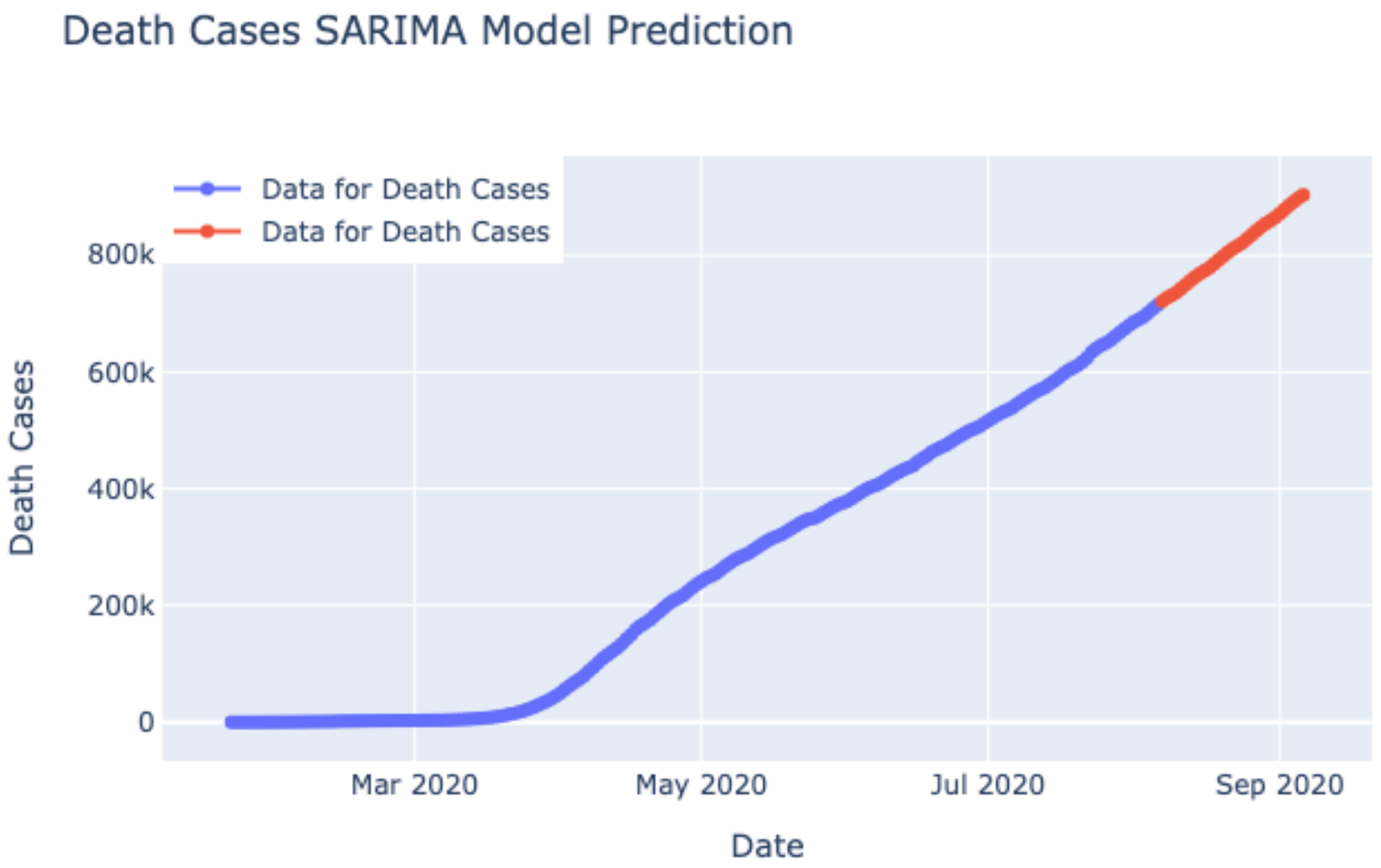
Dep. Variable:	y	**No. Observations:**	188
Model:	SARIMAX(0, 2, 1)x(2, 0, 1, 7)	**Log Likelihood**	-1956.394
Date:	Sat, 22 Aug 2020	**AIC**	3922.787
Time:	15:46:53	**BIC**	3938.916
Sample:	0	**HQIC**	3929.323
	- 188		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.6377	0.057	-11.152	0.000	-0.750	-0.526
ar.S.L7	1.1265	0.119	9.438	0.000	0.893	1.360
ar.S.L14	-0.1280	0.119	-1.077	0.281	-0.361	0.105
ma.S.L7	-0.6826	0.087	-7.852	0.000	-0.853	-0.512
sigma2	7.097e+07	3.13e-09	2.27e+16	0.000	7.1e+07	7.1e+07

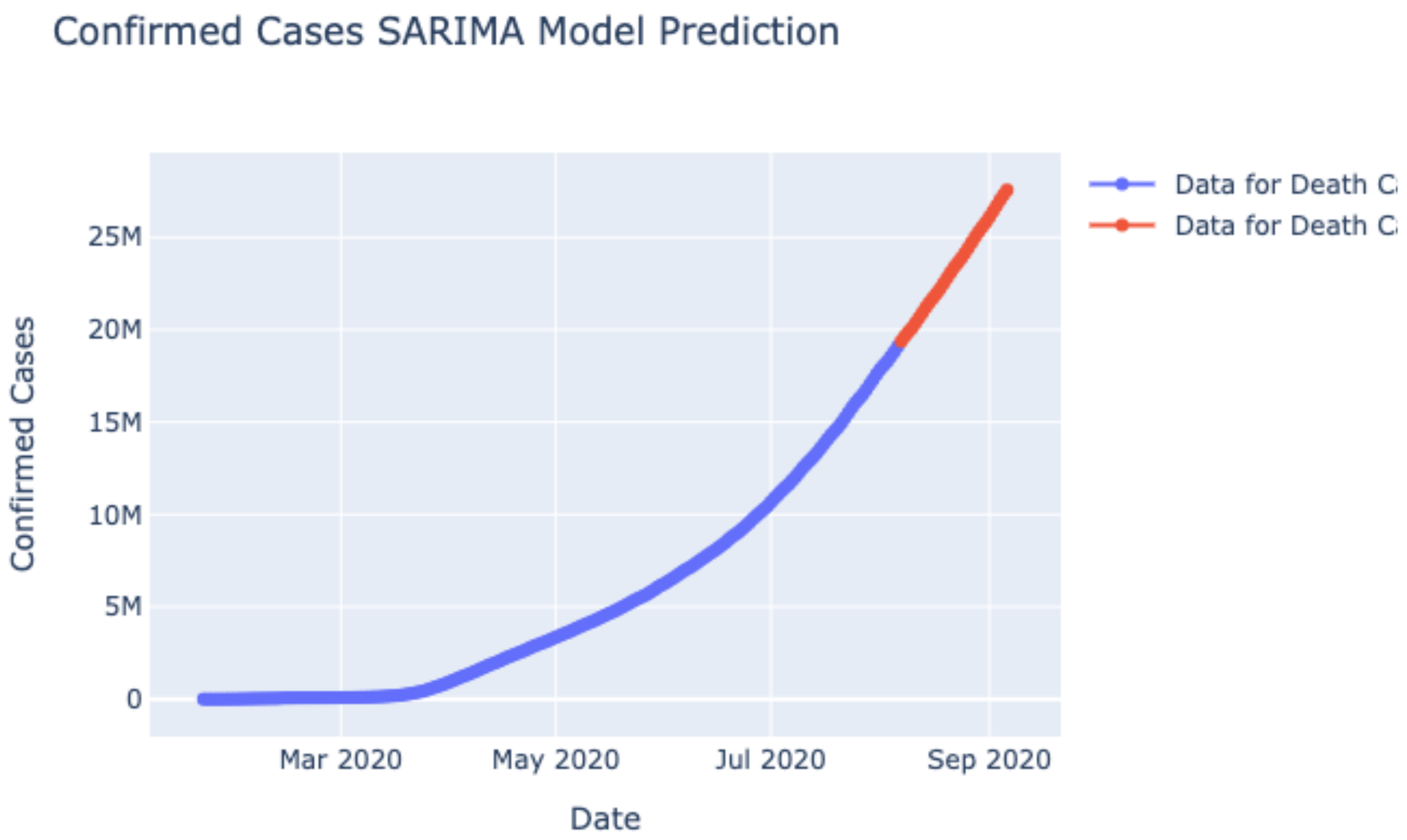
Ljung-Box (Q):	109.94	**Jarque-Bera (JB):**	167.62
Prob(Q):	0.00	**Prob(JB):**	0.00
Heteroskedasticity (H):	9.37	**Skew:**	0.83
Prob(H) (two-sided):	0.00	**Kurtosis:**	7.34

SOME DETAILS ON BEST MODEL(SARIMA)





FORECASTING 30 DAYS DEATHS CASES



FORECASTING 30 DAYS CONFIRMED CASES

Assumptions and Limitations

- Stationarity: The first **assumption** is that the **series** are stationary.
- This means that the series are normally distributed and the mean and variance are constant over a long time period.
- In Univariate Time Series analysis, exogeneous factors are not taken consideration due to which forecasting may differ if considered those factors.

More Ideas to improve model in future

- In this case , only one variable is observed at each time is called ‘Univariate Time Series’.
- If two or more variables are observed at each time is called ‘Multivariate Time Series’ . In future I would consider exogenous factor to forecast using Multivariate Time series models.
- In this case, we will focus on the univariate time series for forecasting the cases with

  Auto SARIMA functionality in python.
- I will use Multivariate Time series models to forecast cases using LSTM RNN for better results with more data.

Conclusions

- All sources of datasets helps in forecasting of covid-19 cases .

- Out of 4 models , SARIMA model performs best with least 64293.45 and AIC score 3922 scores.
- Model has forecasted increase of death cases to 903348 and confirmed cases to 27,564,467 by 2020-09-06 worldwide.