

# Covid-19 Analysis and Forecasting

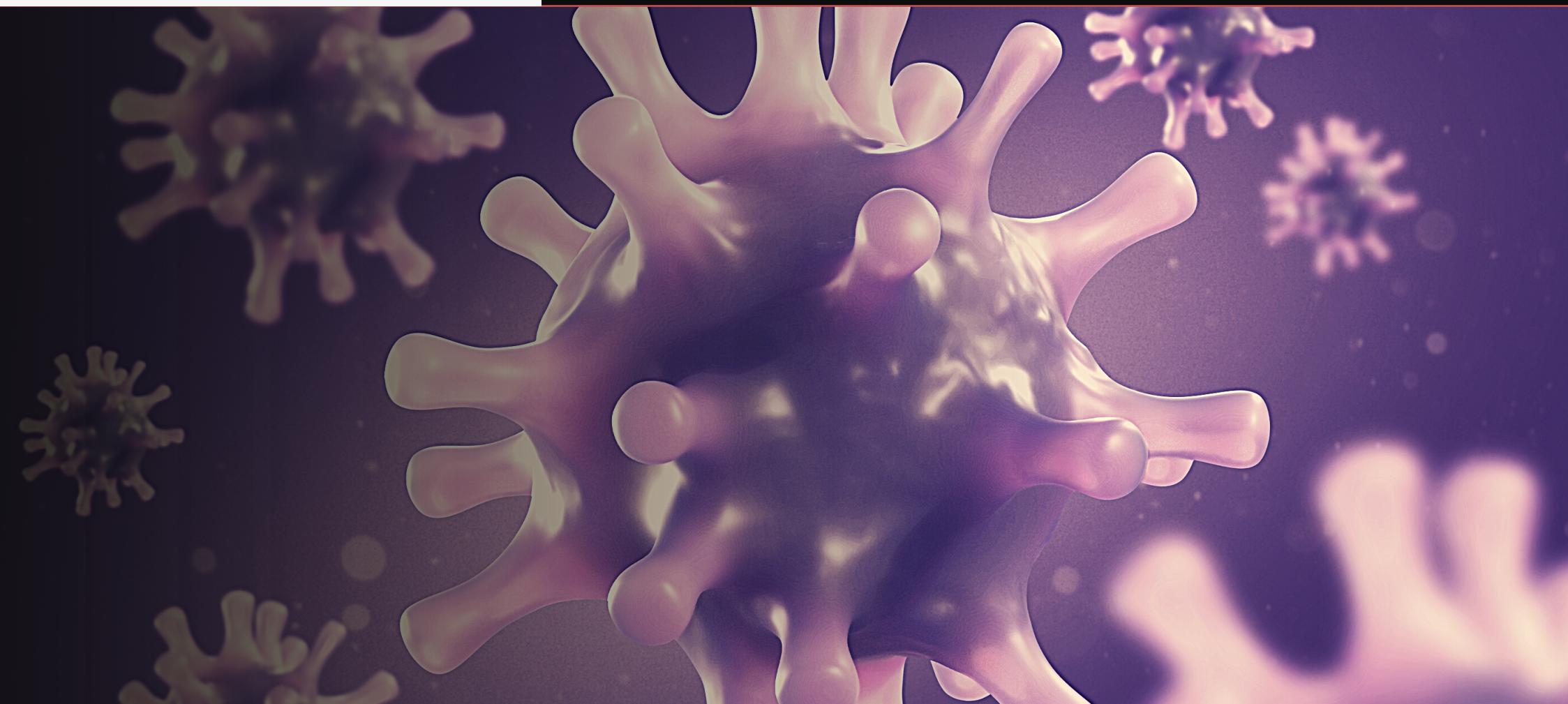
## Reeti Bhagat

Data Science intensive capstone project May 26th, 2020 Cohort



---

Thanks to Springboard mentor  
Ash Yousefi



## 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?

---

- Covid-19 spreads at an alarmingly exponential rate, unlike other diseases.
- Insufficient time for proper planning in terms of medical supplies and other measures to stabilize the death rate.
- No definite line of treatment.

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
- Casino and gaming
- 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/](https://www.worldometers.info/coronavirus/)

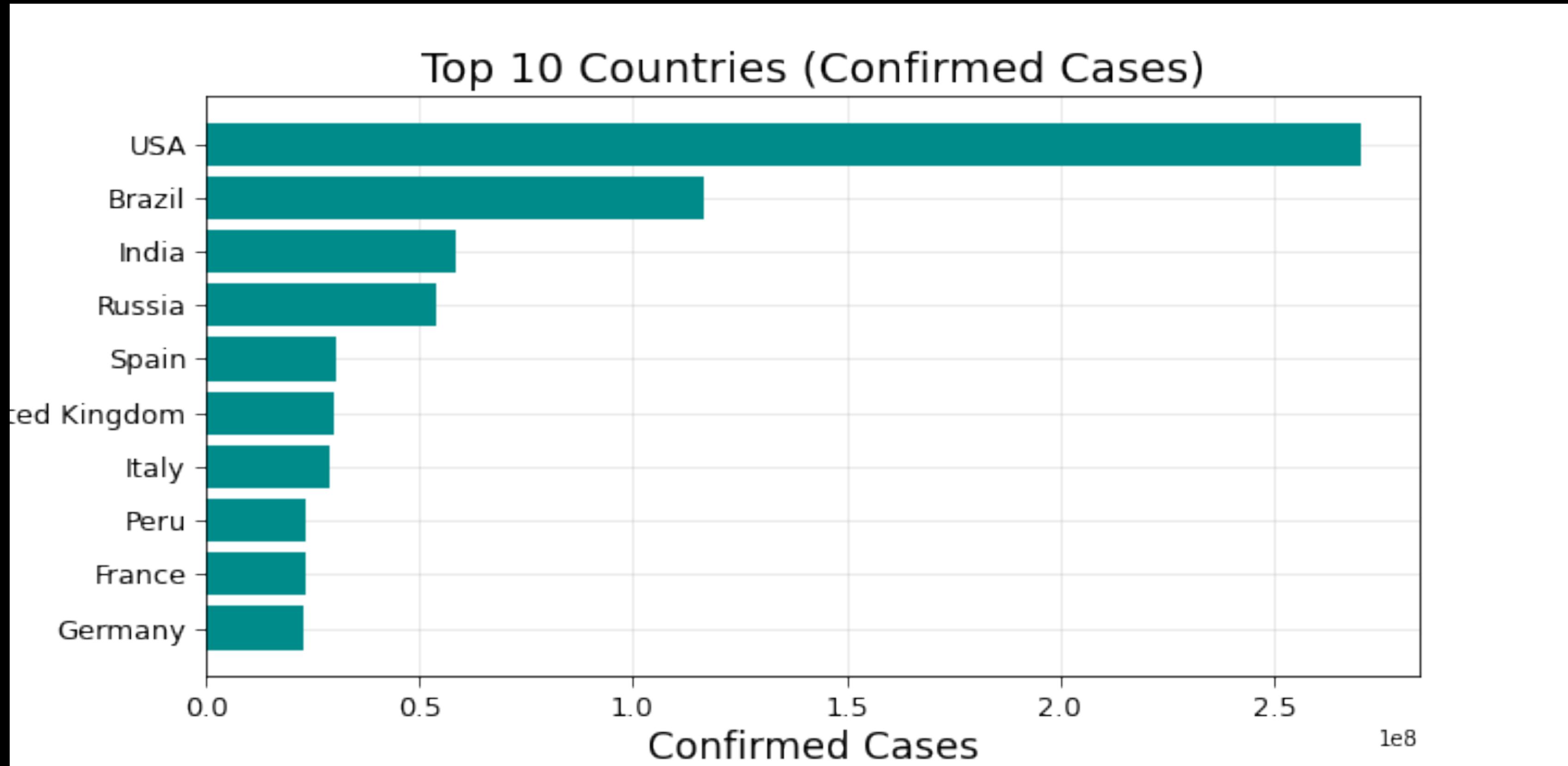
# Data Exploration

[https://github.com/reetibhagat/capstone-1-covid-19/blob/master/notebooks/Exploratory data analysis capstone 1.ipynb](https://github.com/reetibhagat/capstone-1-covid-19/blob/master/notebooks/Exploratory%20data%20analysis%20capstone%20%201.ipynb)

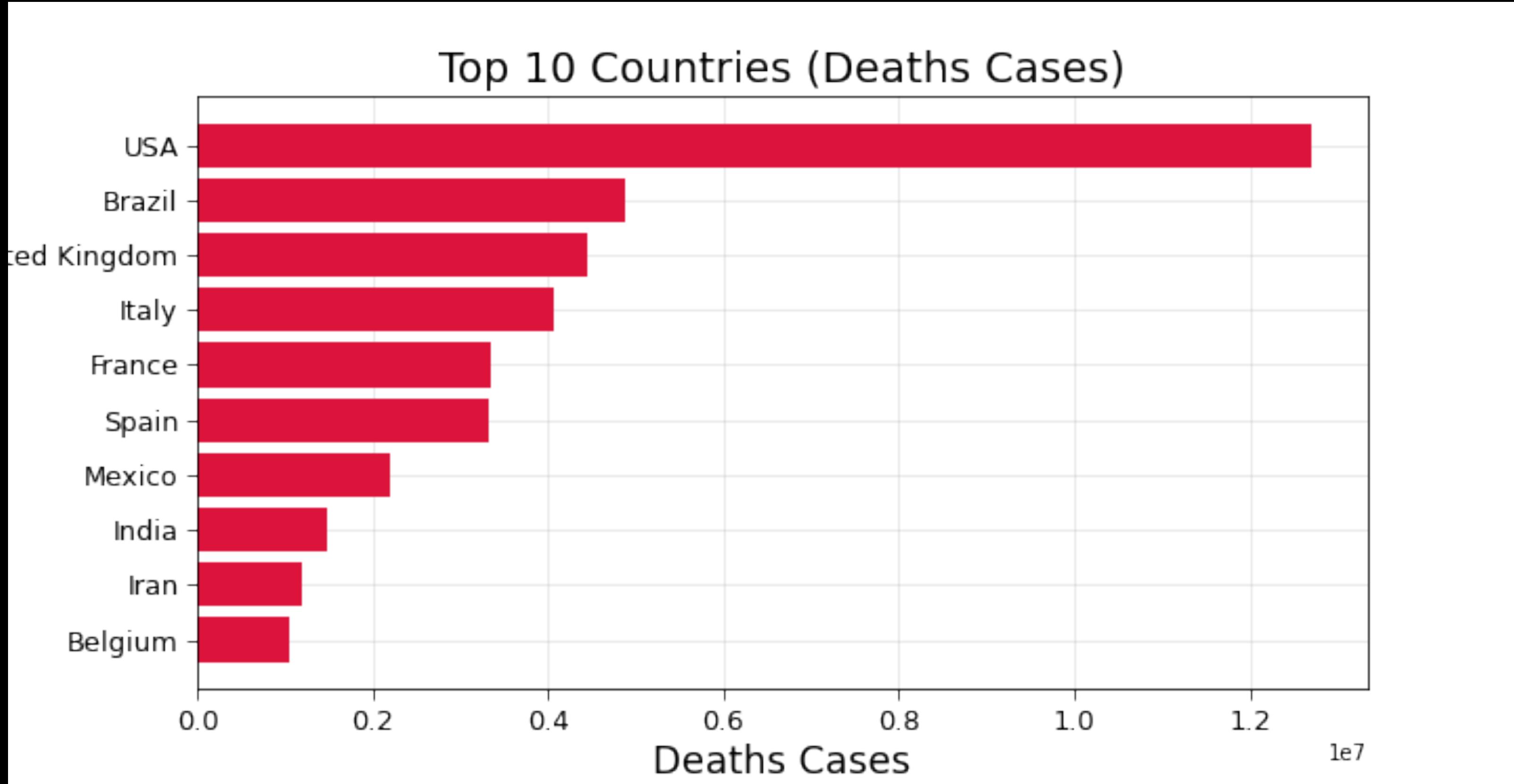
## General Analysis of Data( Continent wise Data)

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

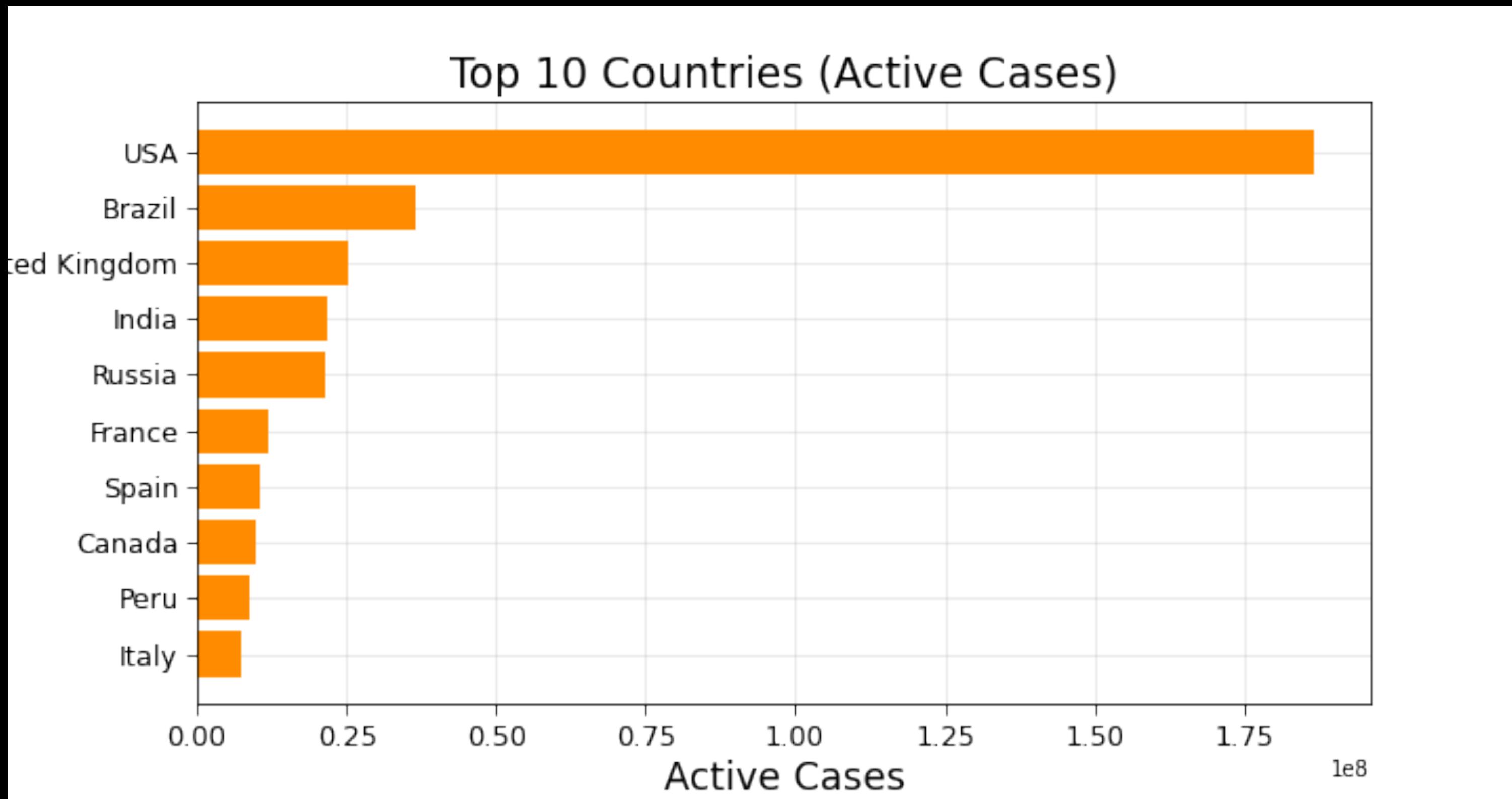
# Top 10 COUNTRIES(CONFIRMED CASES)



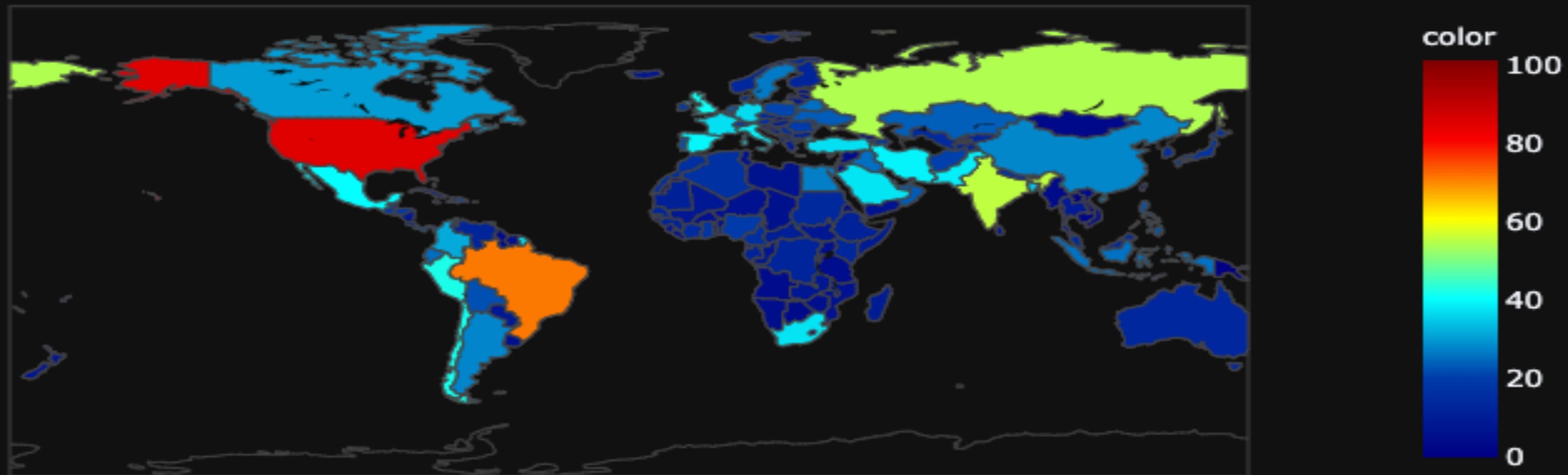
# Top 10 Countries(Deaths Cases)



# Top 10 Countries(Active Cases)



# Covid-19:Progression of Spread

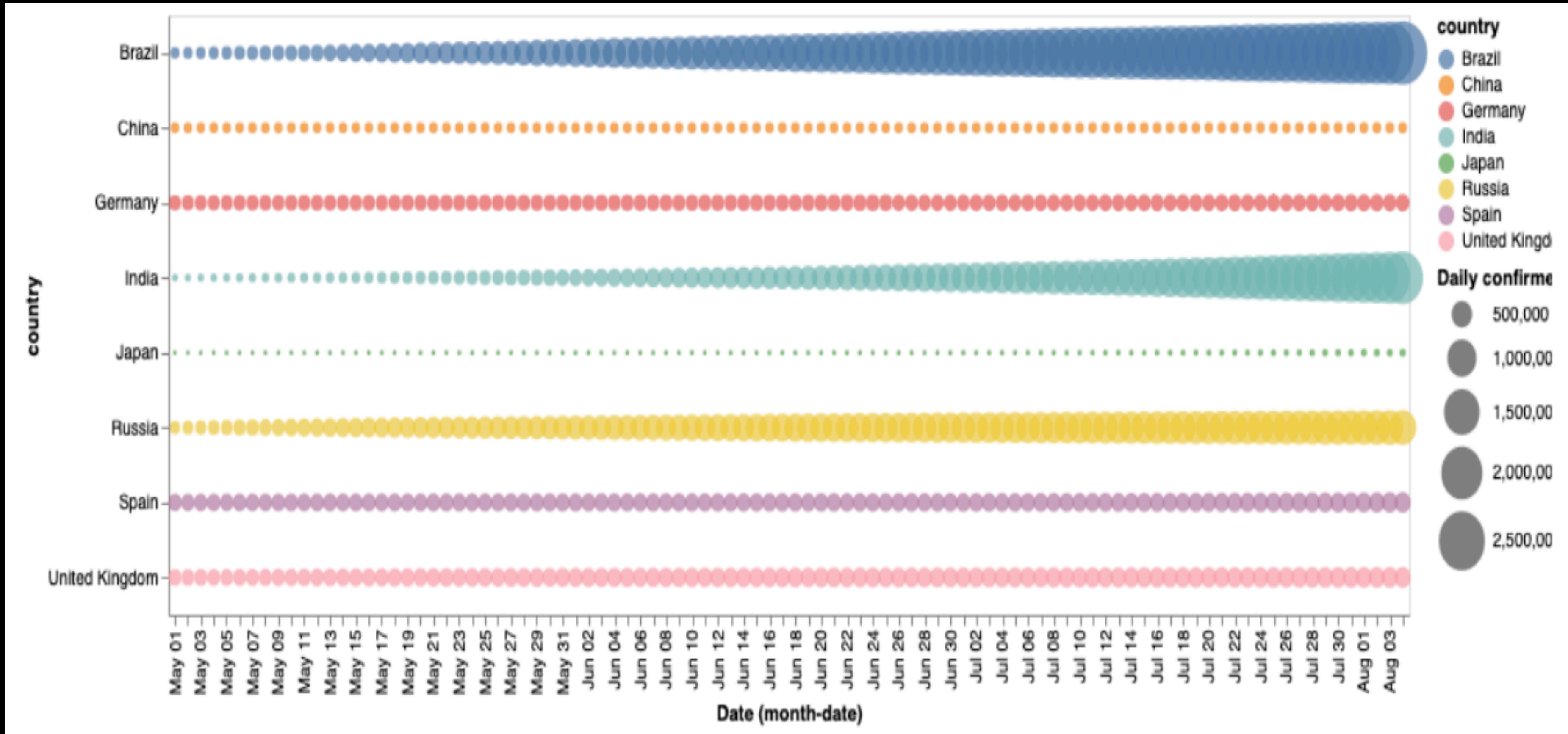


Date=07/07/2020



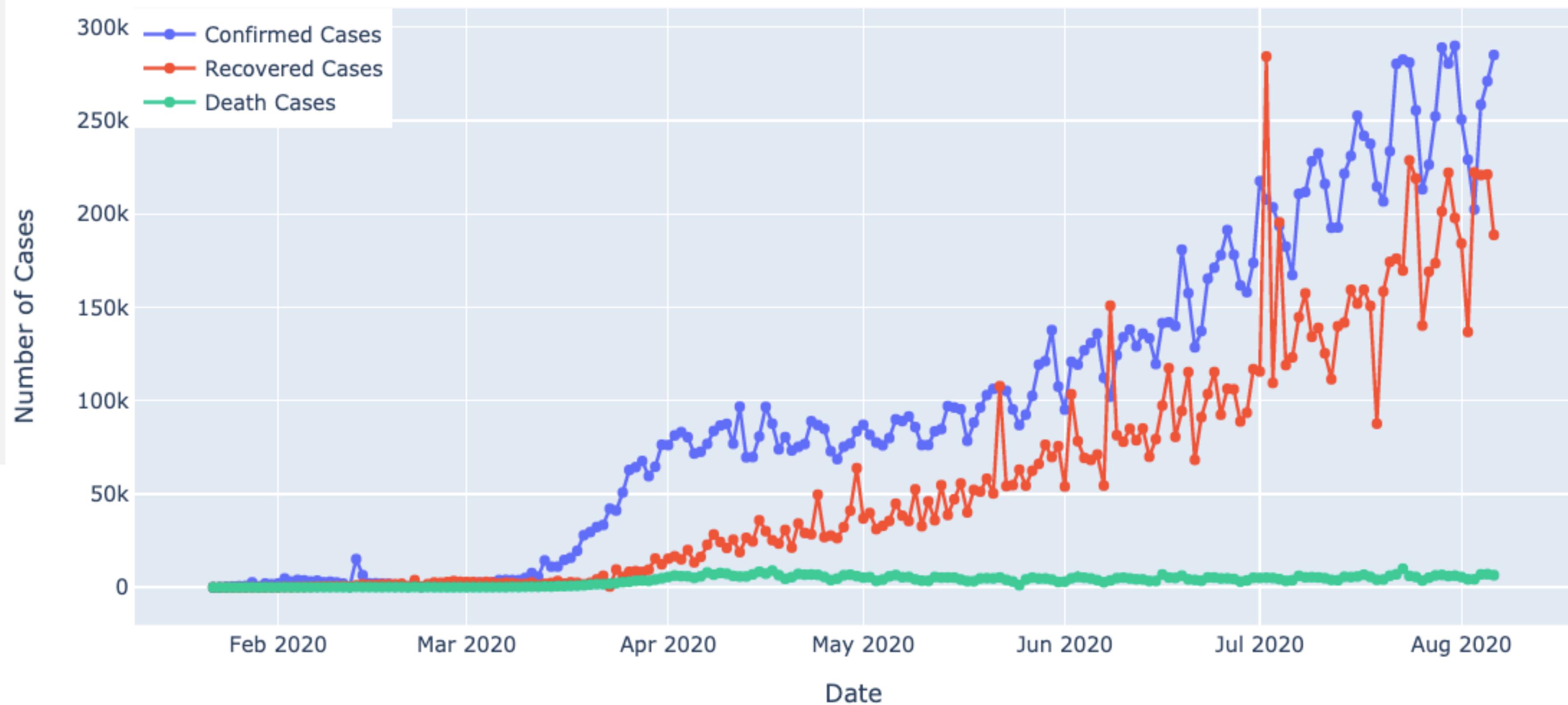
01/22/2020 02/22/2020 03/24/2020 04/24/2020 05/25/2020 06/25/2020 07/26/2020

# Comparison of spread of Covid-19 in different countries

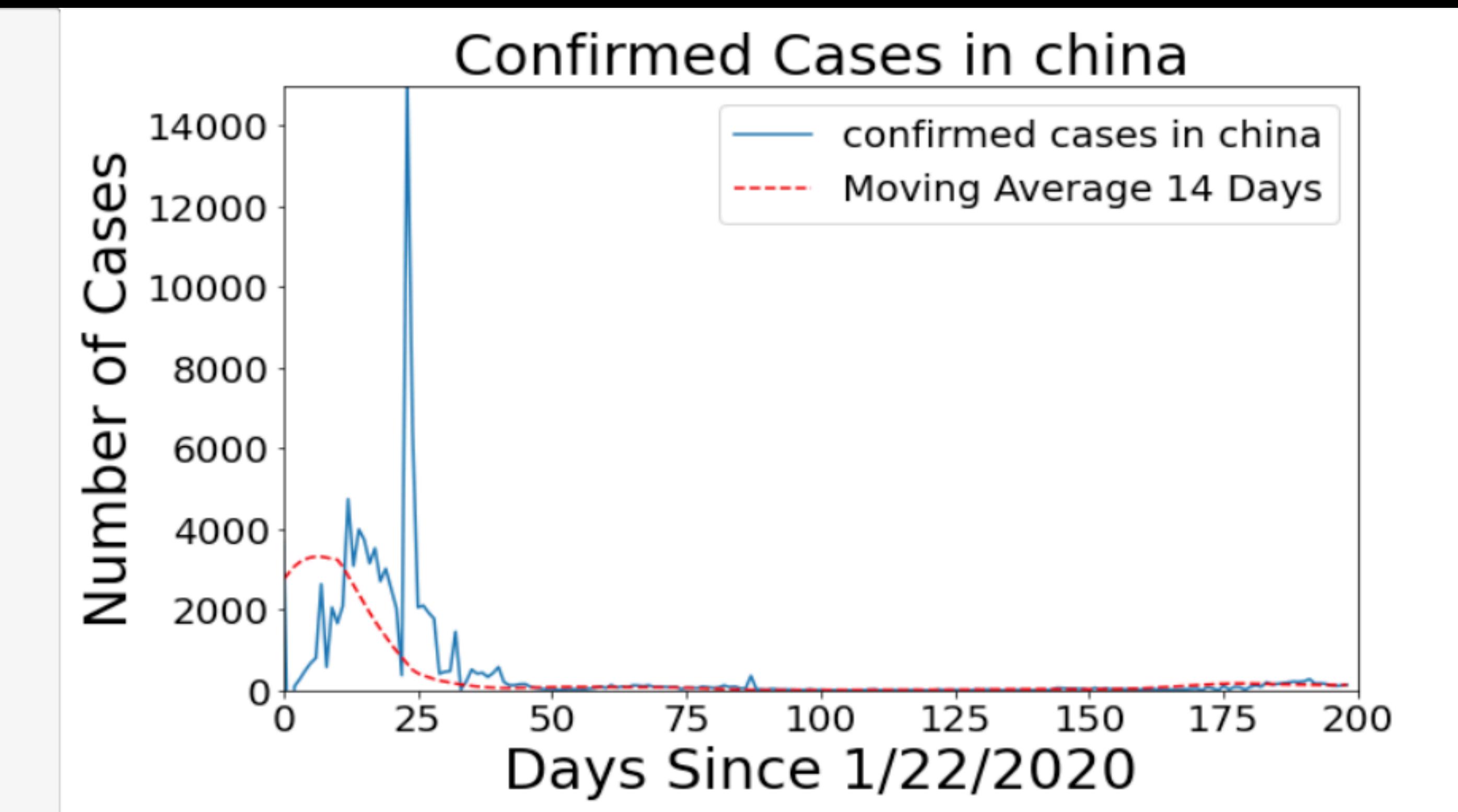




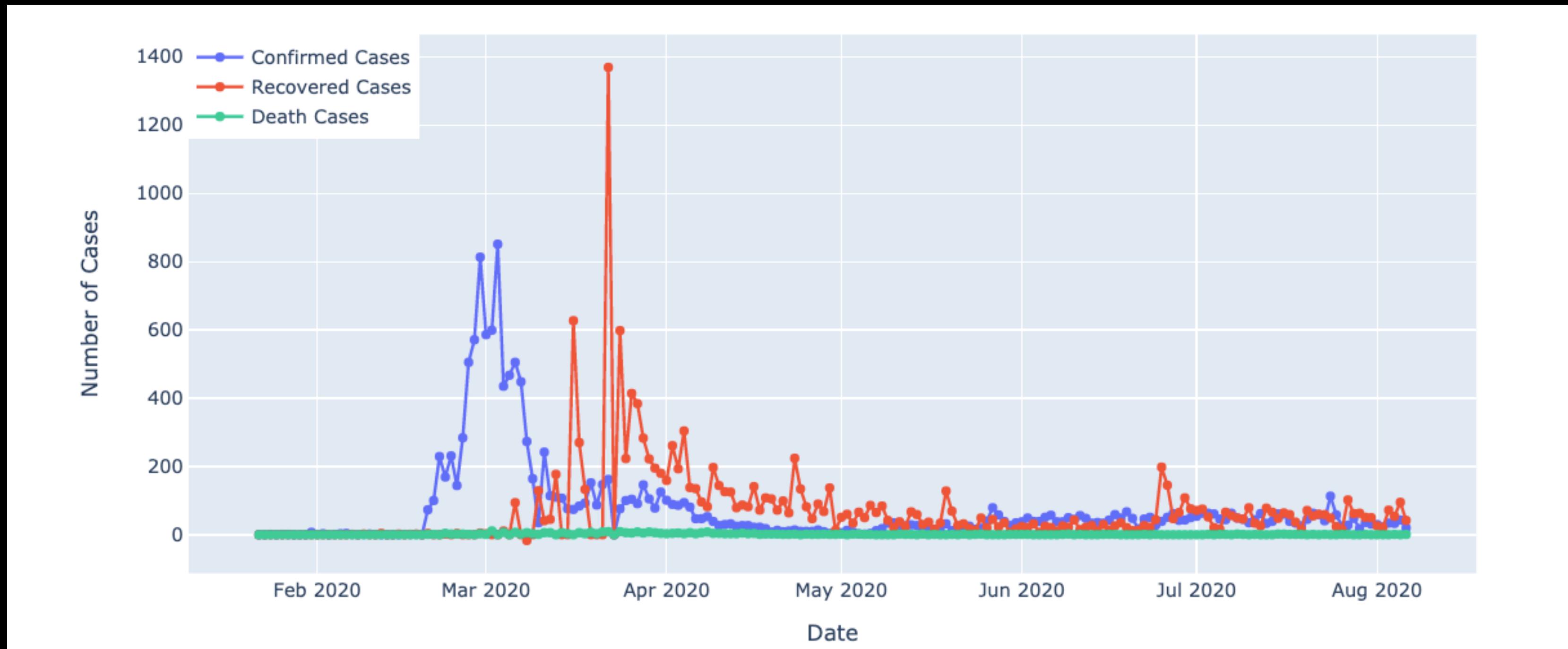
## Daily increase in different types of Cases



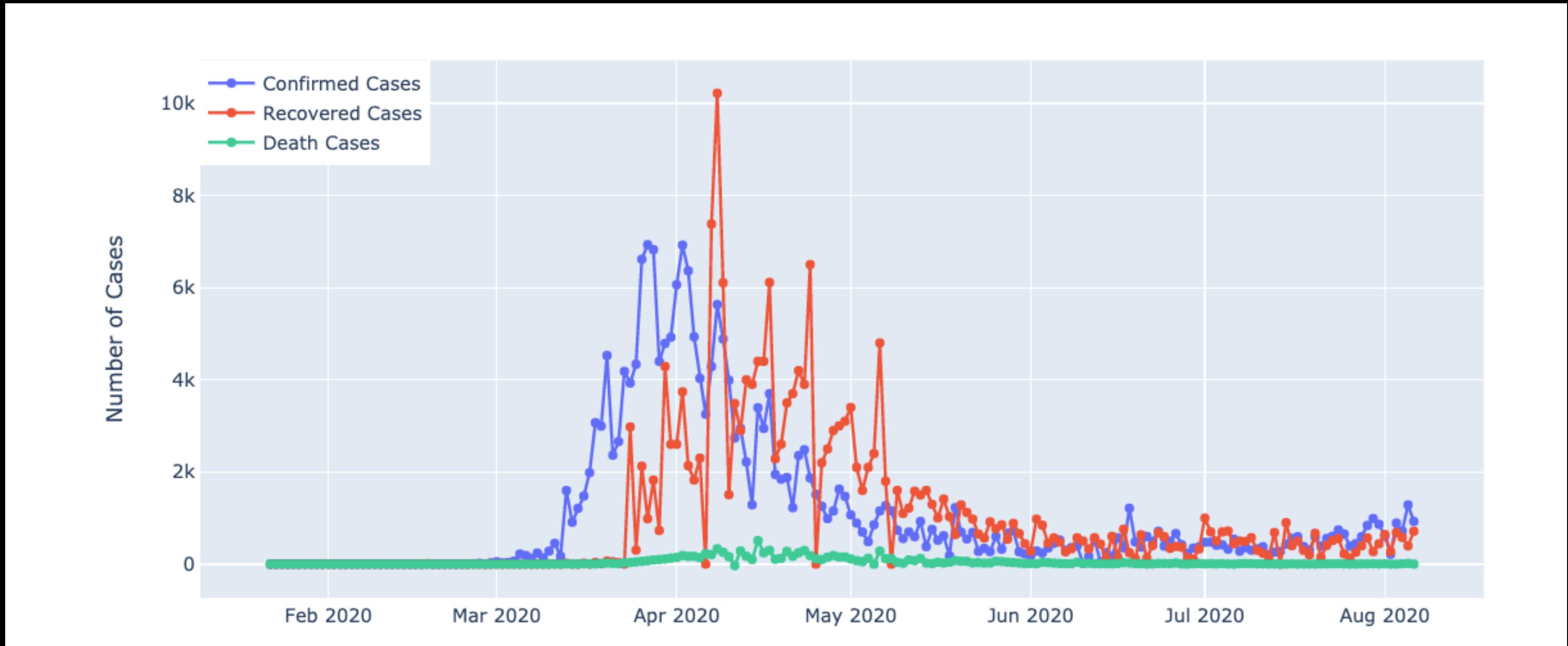
# Confirmed Cases in China



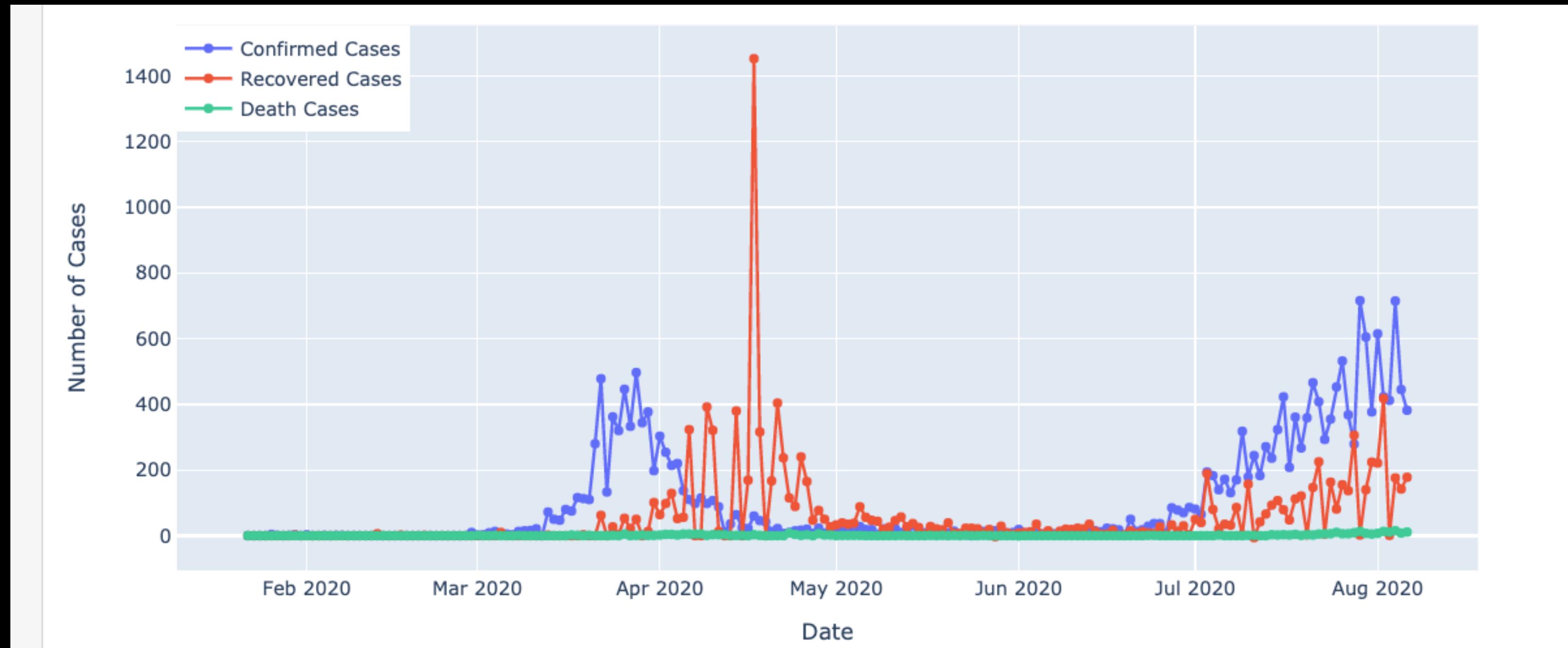
# South Korea



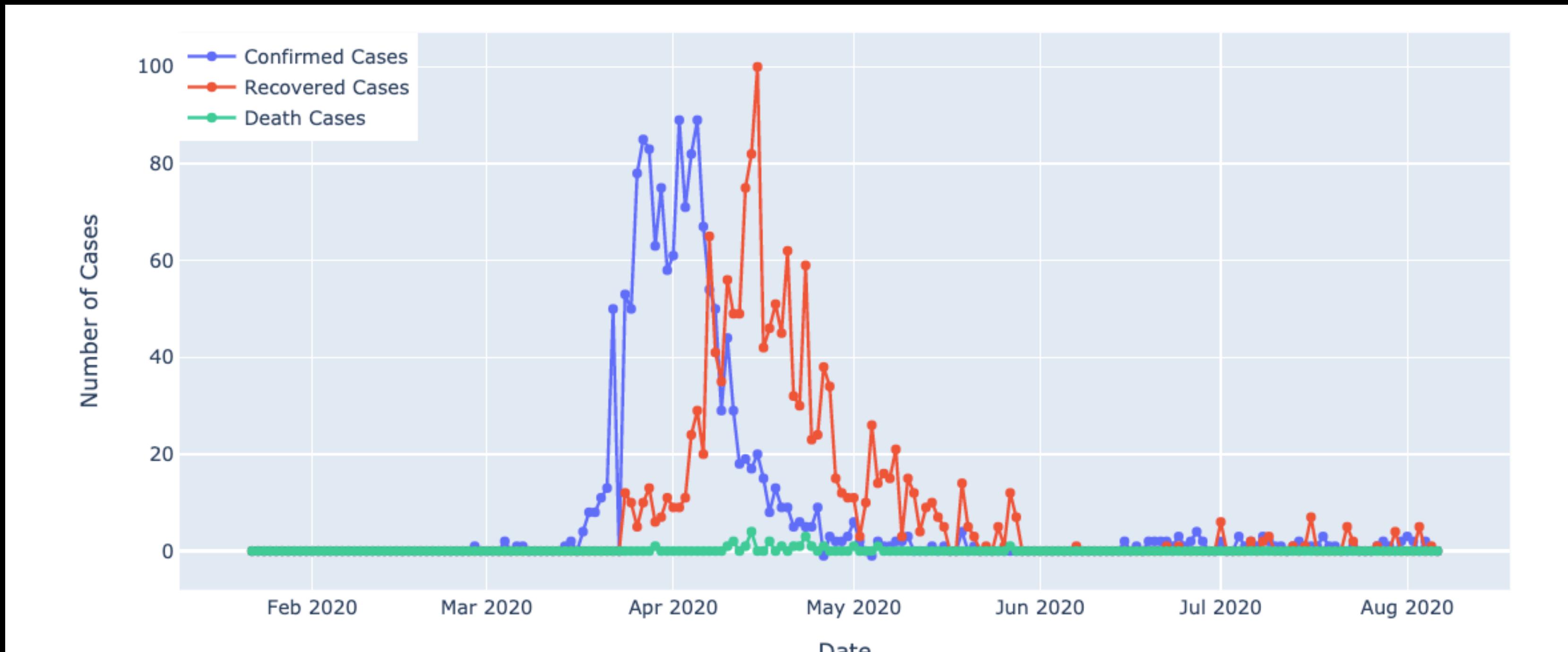
# Germany

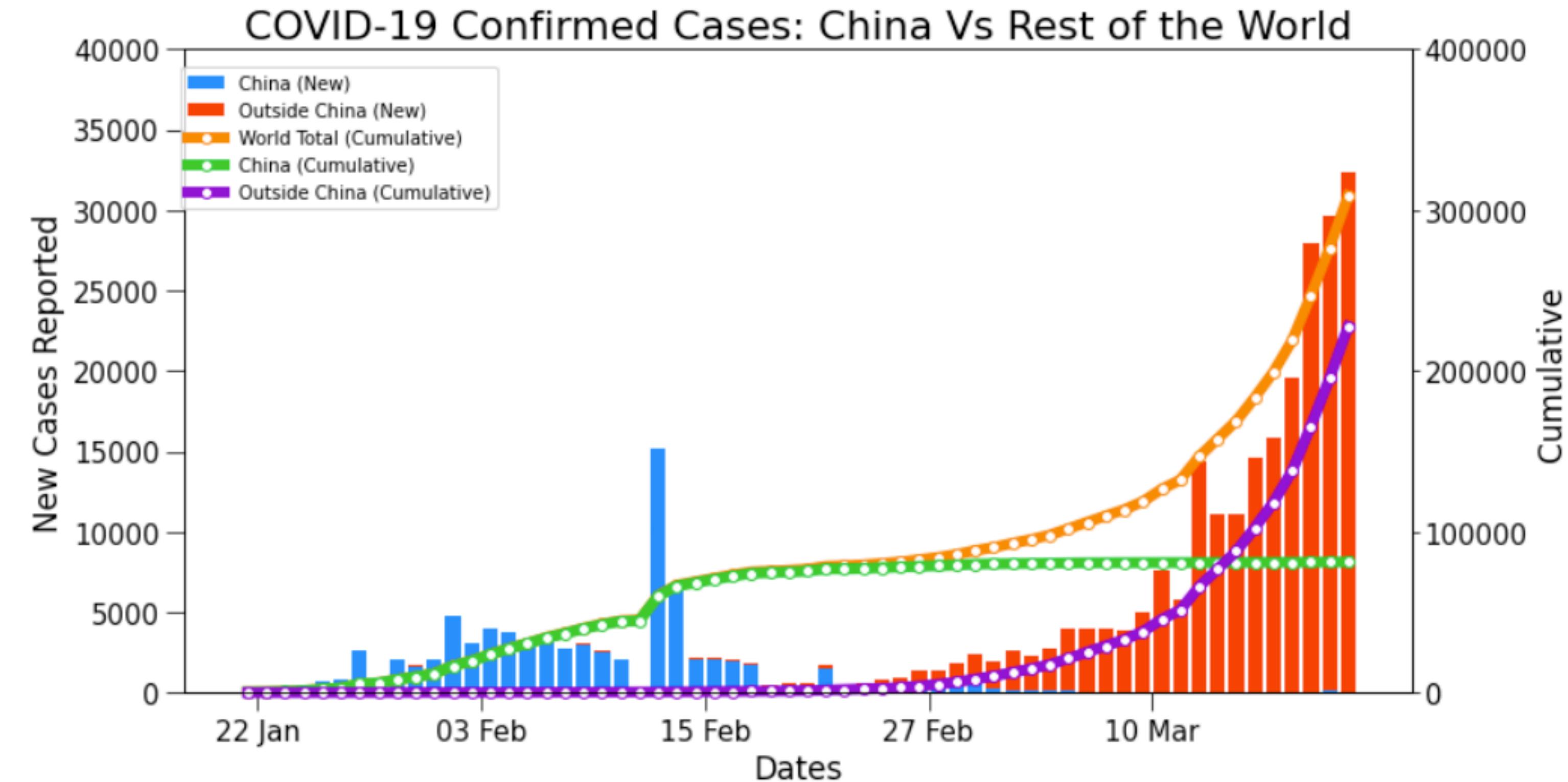


# Australia

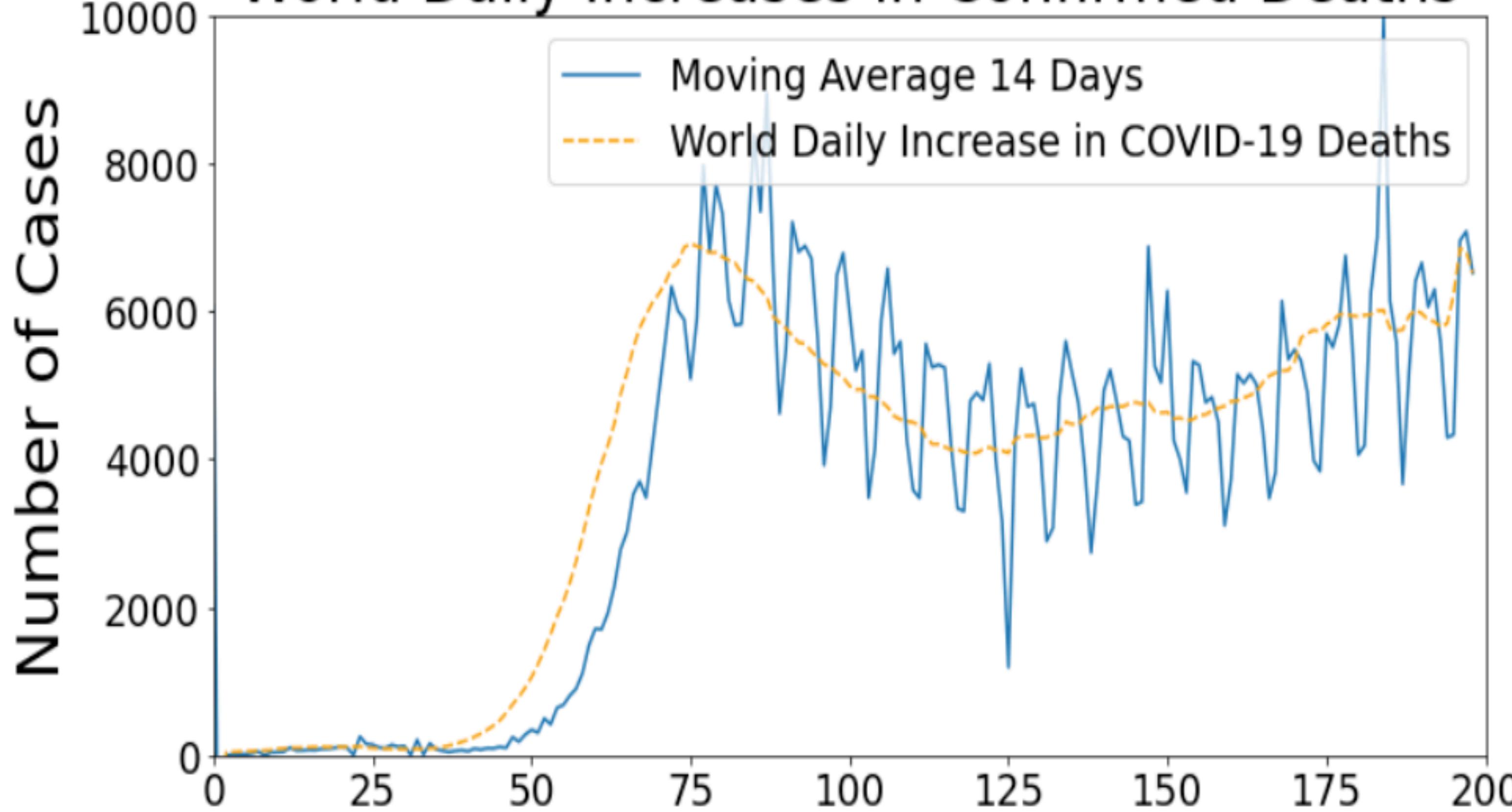


# New Zealand



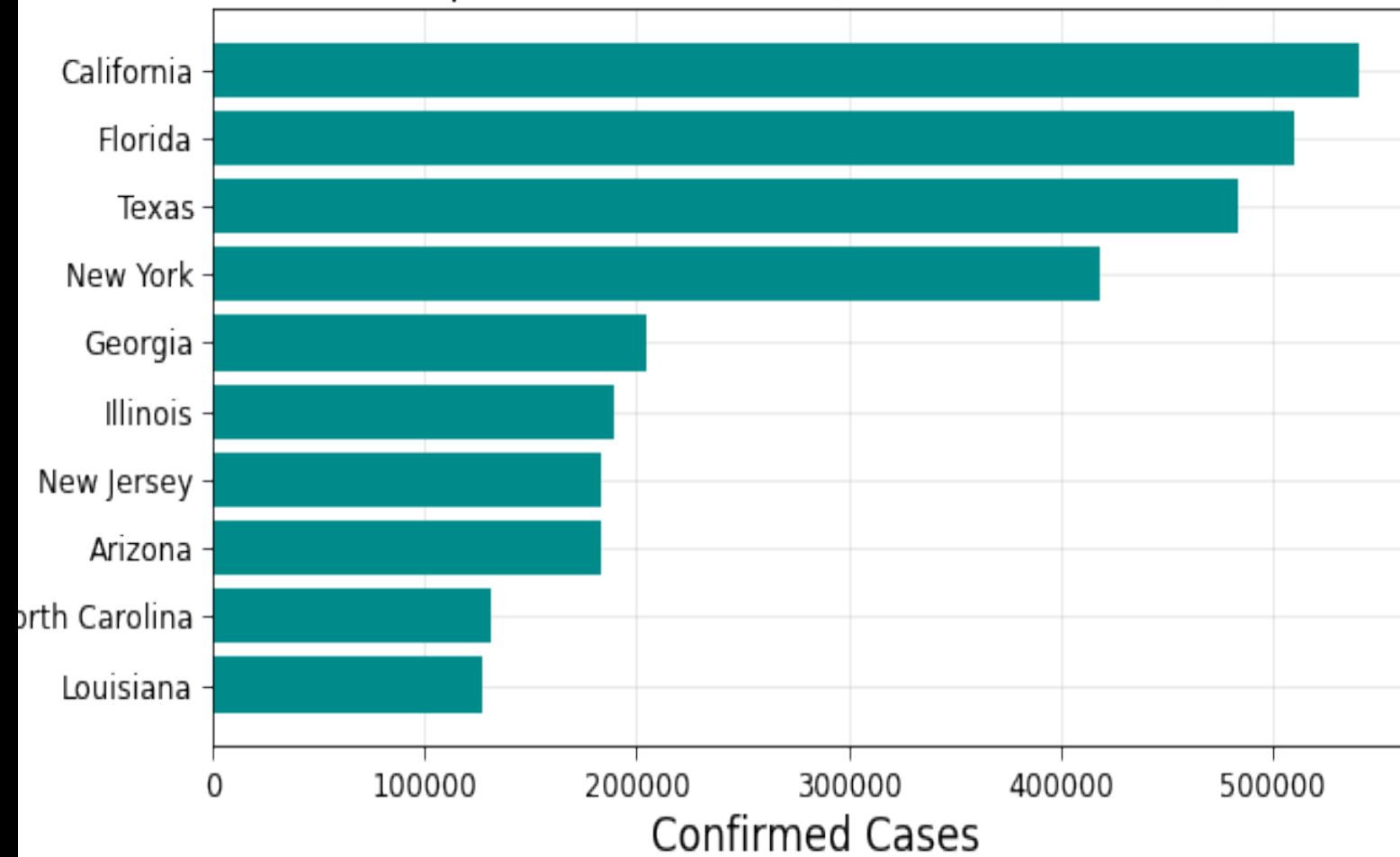


# World Daily Increases in Confirmed Deaths

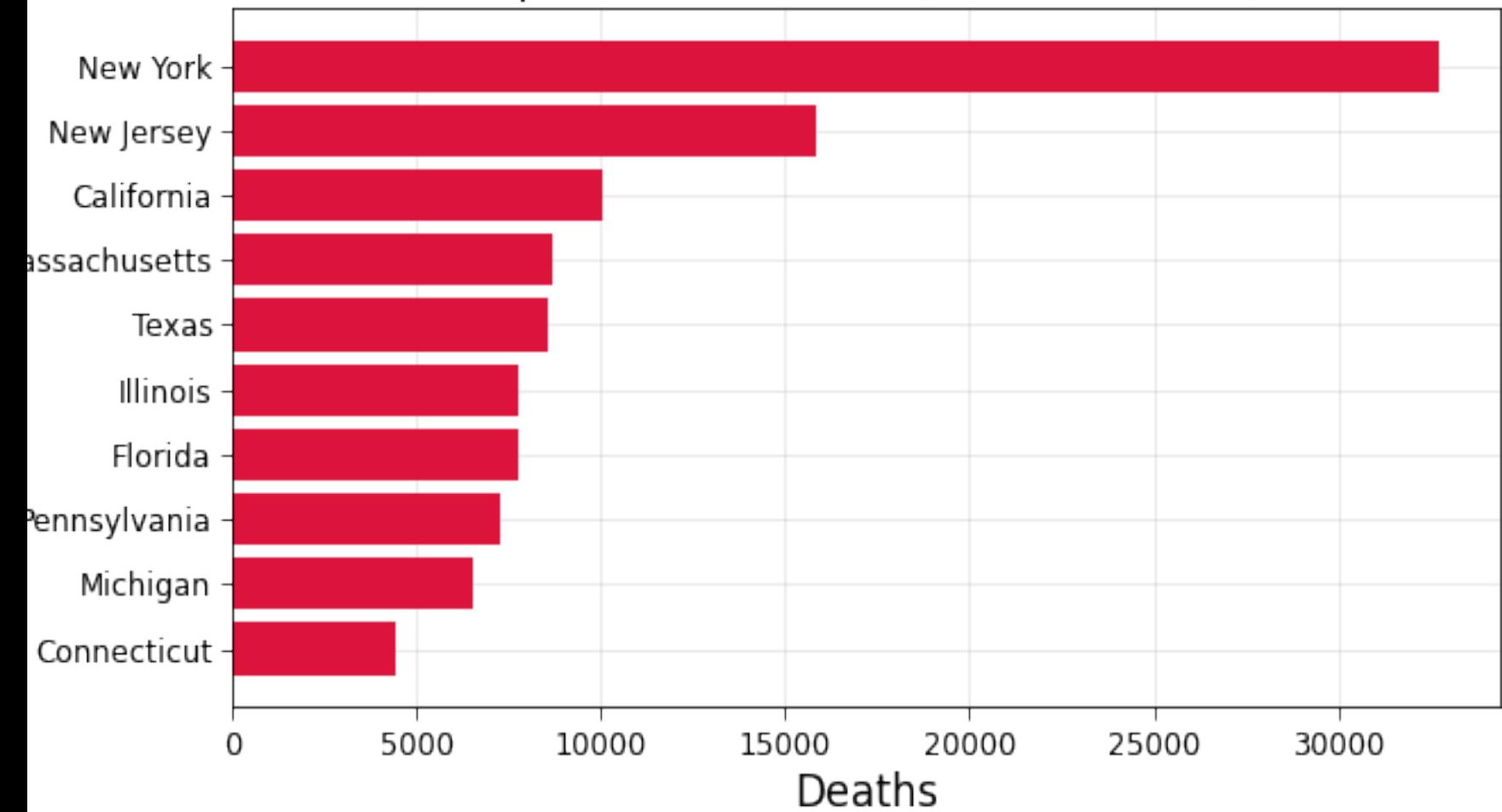


# USA

Top 10 States: USA (Confirmed Cases)



Top 10 States: USA (Deaths Cases)



# Machine learning Modeling

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

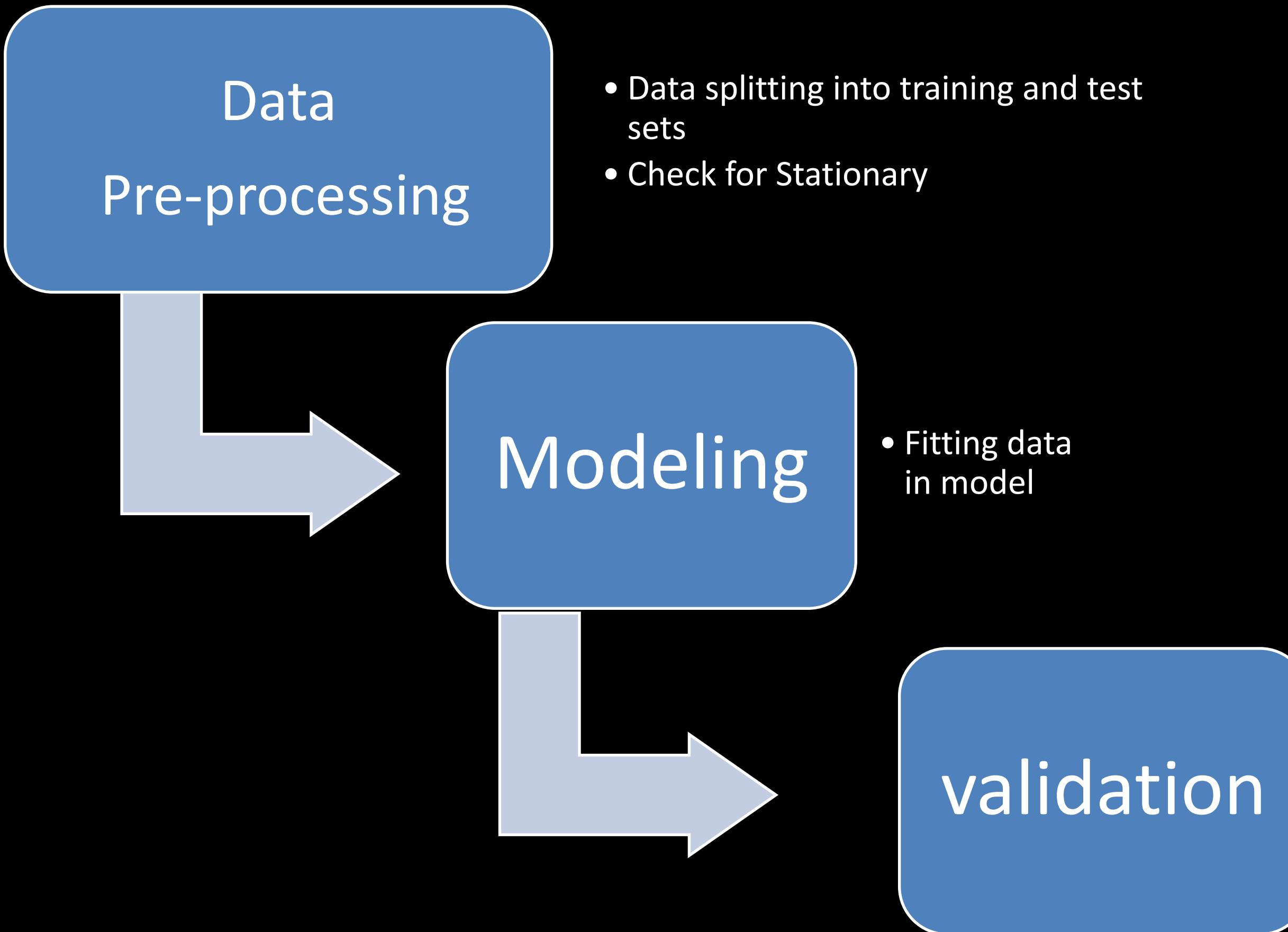
Time Series Analysis

# 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



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

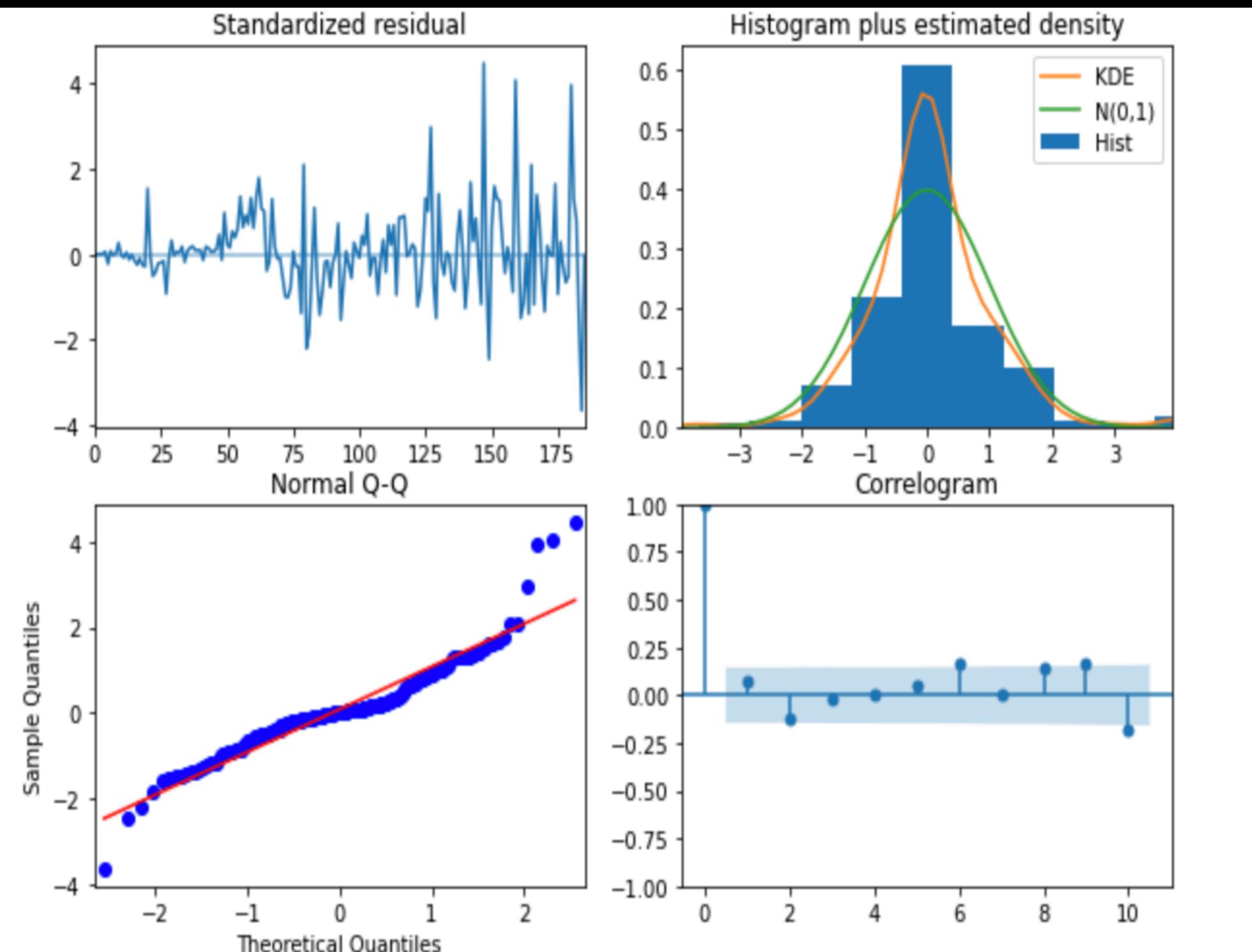
# Models Comparison

- 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.

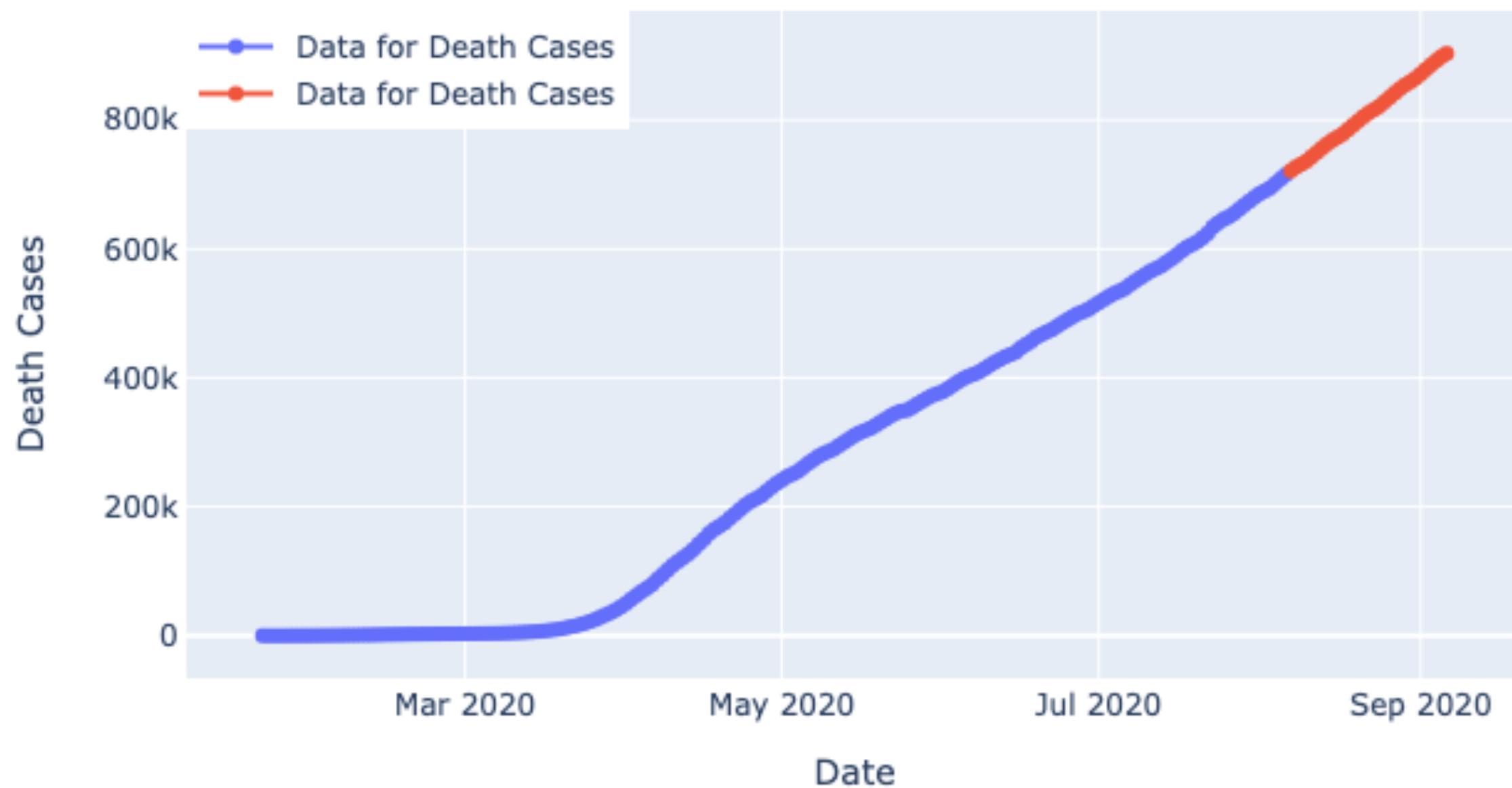
# SOME DETAILS ON BEST MODEL(SARIMA)

SARIMAX Results

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

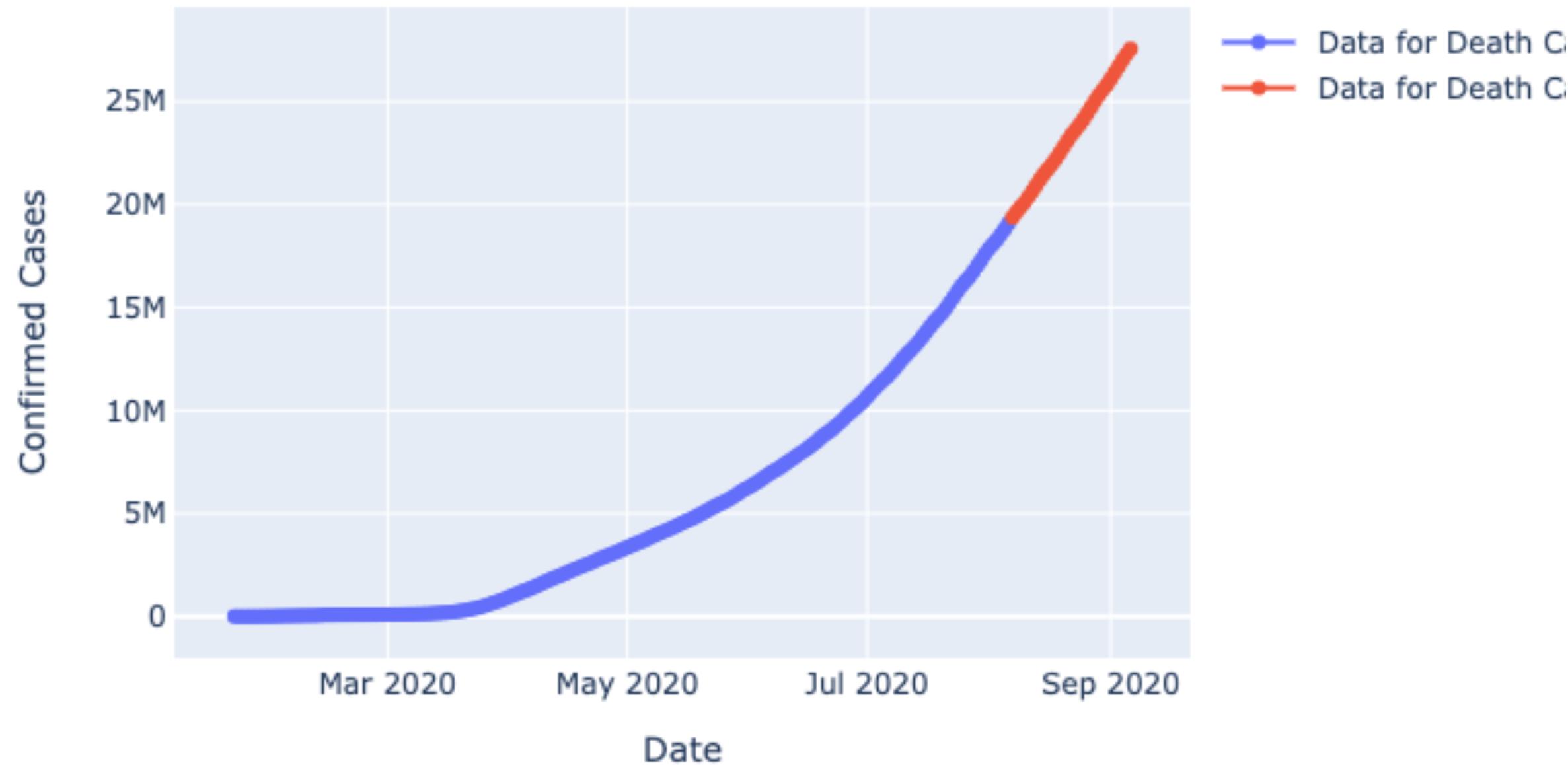


## Death Cases SARIMA Model Prediction



FORECASTING  
30 DAYS DEATHS  
CASES

## Confirmed Cases SARIMA Model Prediction



FORECASTING  
30 DAYS  
CONFIRMED  
CASES

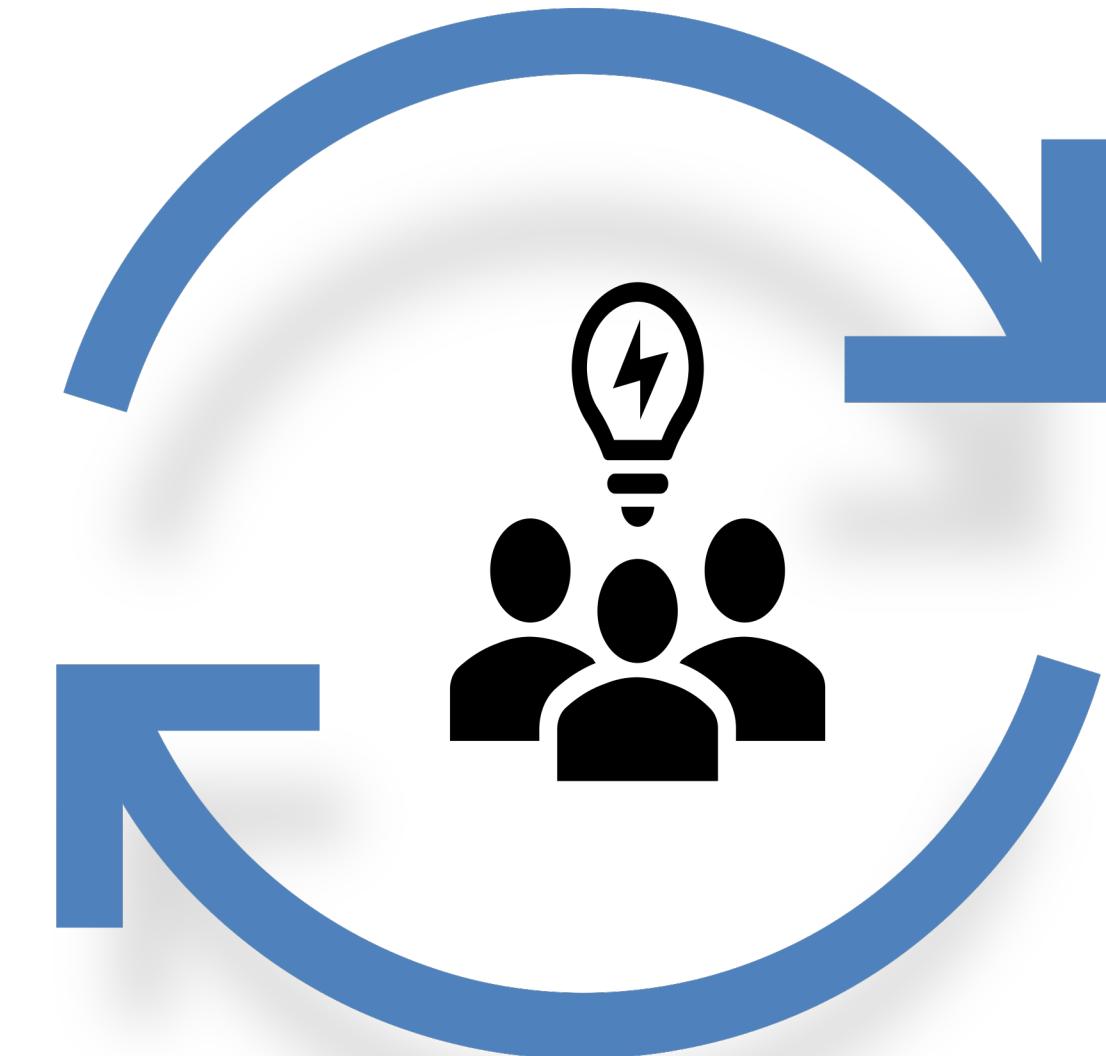
# Assumptions and Limitations



- 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.
- No outliers: We **assume** that there is no outlier in the **series**.
- In Univariate Time Series analysis, exogenous 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 RMSE and AIC scores.
- Model has forecasted increase of death cases to 903348 and confirmed cases to 27,564,467 by 2020-09-06 worldwide.
- With more ideas, model can be improved in future