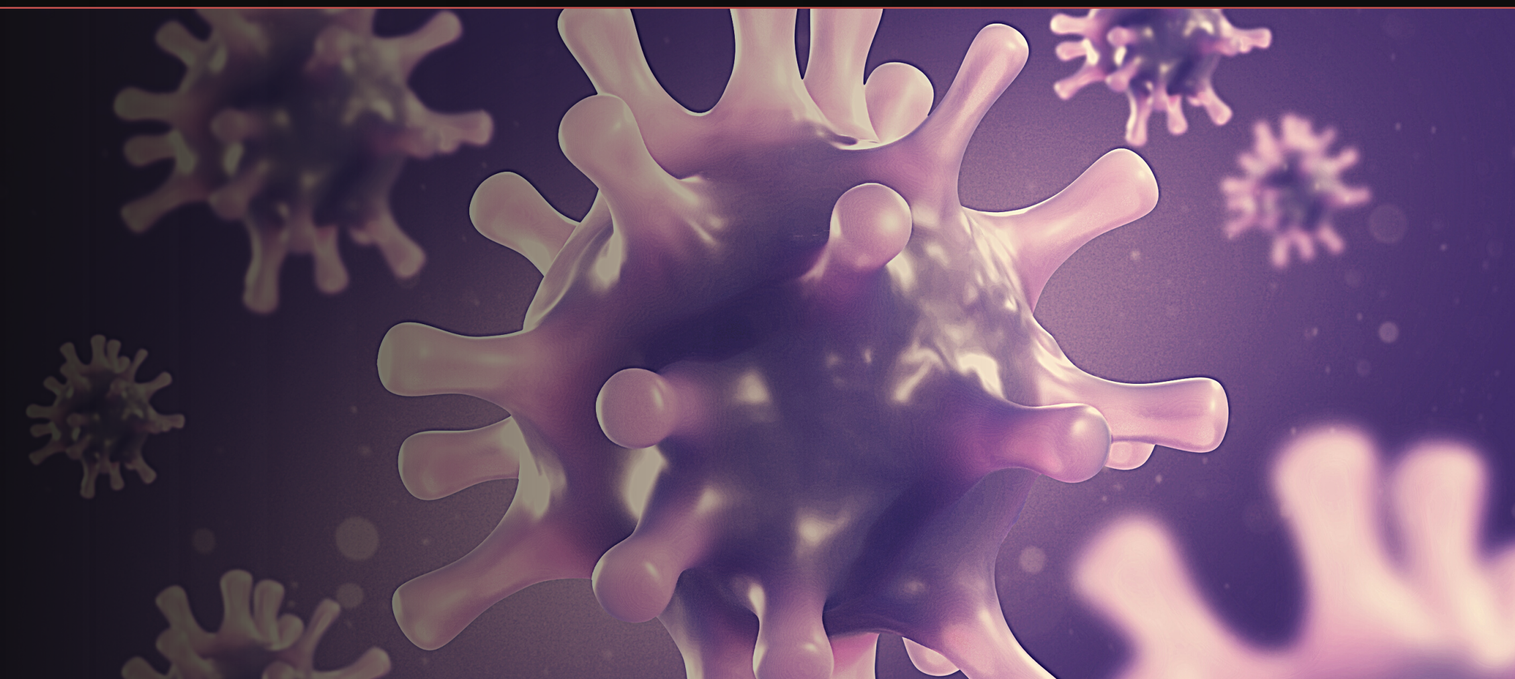
Covid-19 Analysis and Forecasting
Reeti Bhagat

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.

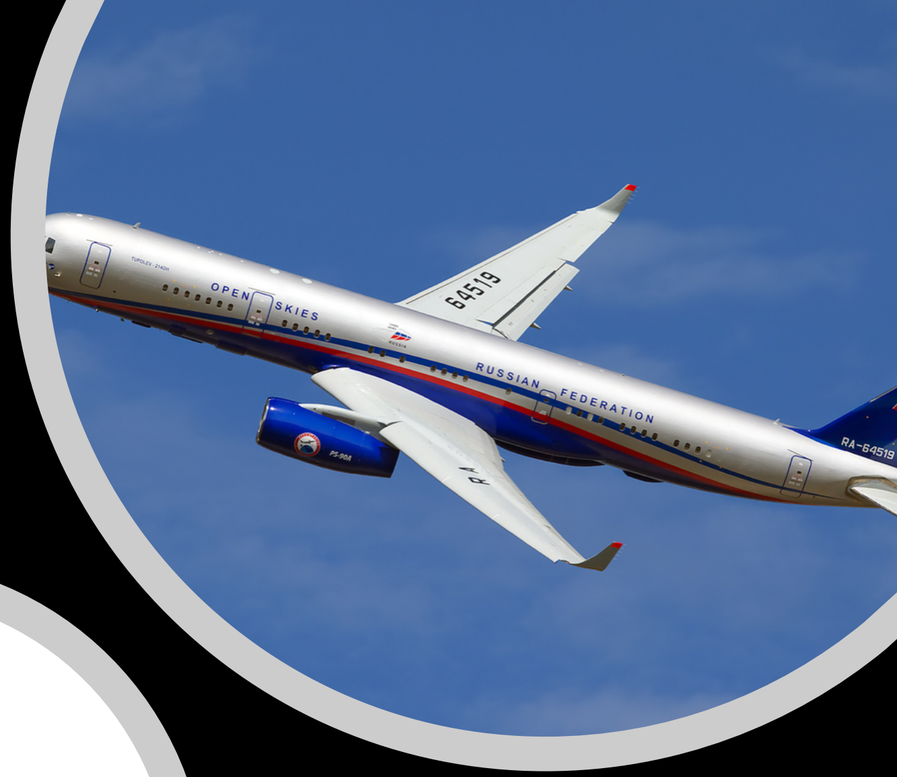
Why should be Concerned?

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/

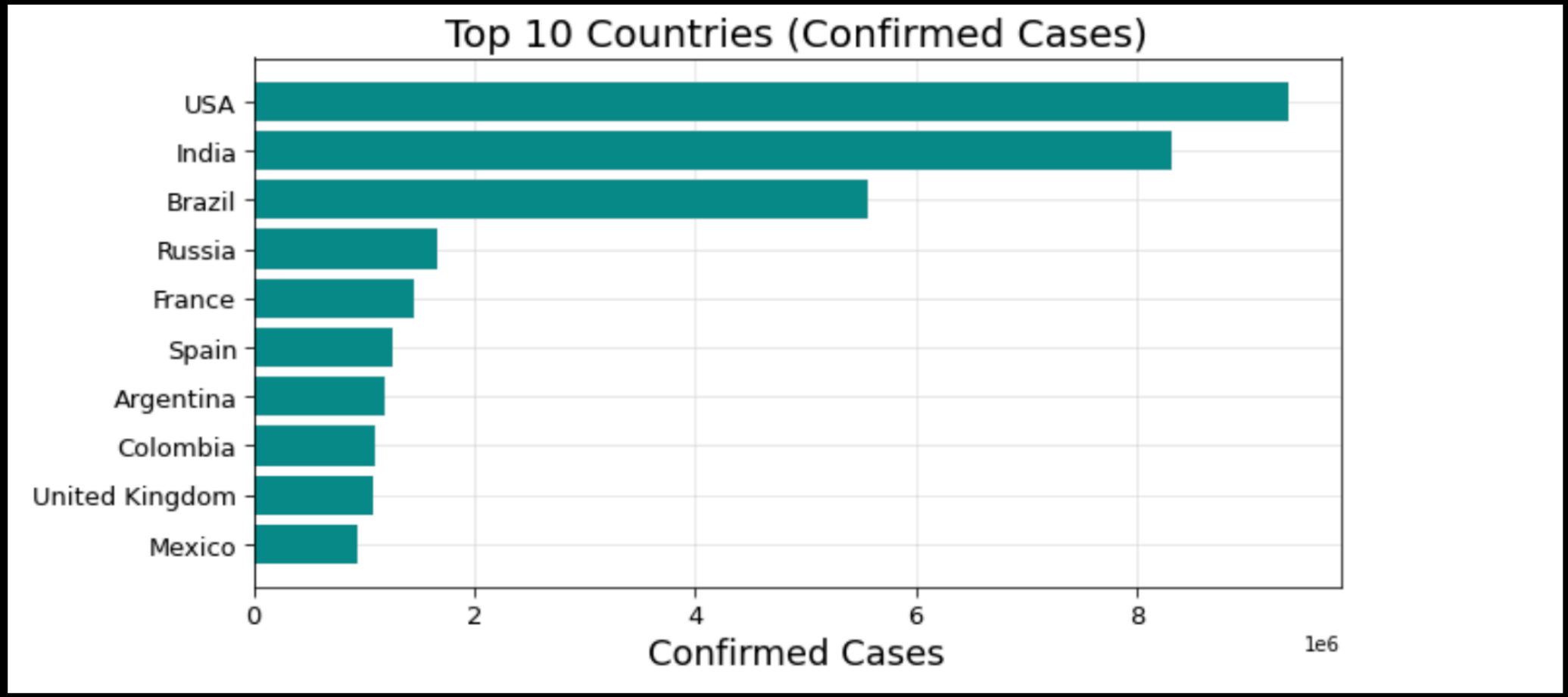
Data Exploration

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

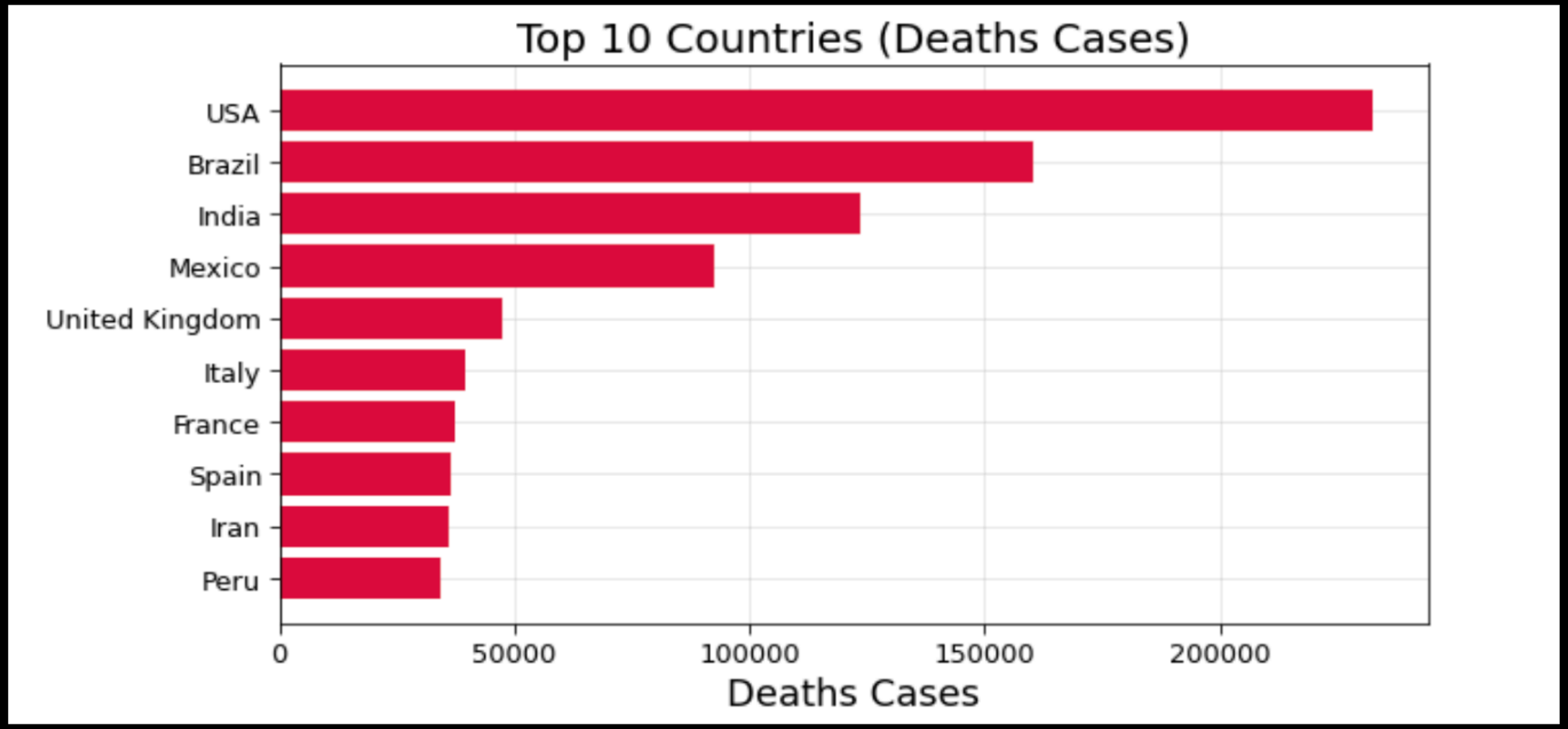
General Analysis of Data(Continent wise Data)

continent	Confirmed	Deaths	Recovered	Mortality Rate	Recovery Rate	Active Cases
Africa	1814269	43705	1470601	116.91	4259.86	299963
Asia	13833712	245578	12344443	92.13	3492.18	1243691
Australia	30239	941	27895	11.61	406.11	1403
Europe	10600198	273334	3649469	82.87	2071.44	6677395
North America	11233556	350451	5001513	51.18	1658.15	5881592
Others	133566	2536	103732	41.89	551.04	27298
South America	9759855	297190	8790456	38.85	1044.52	672209

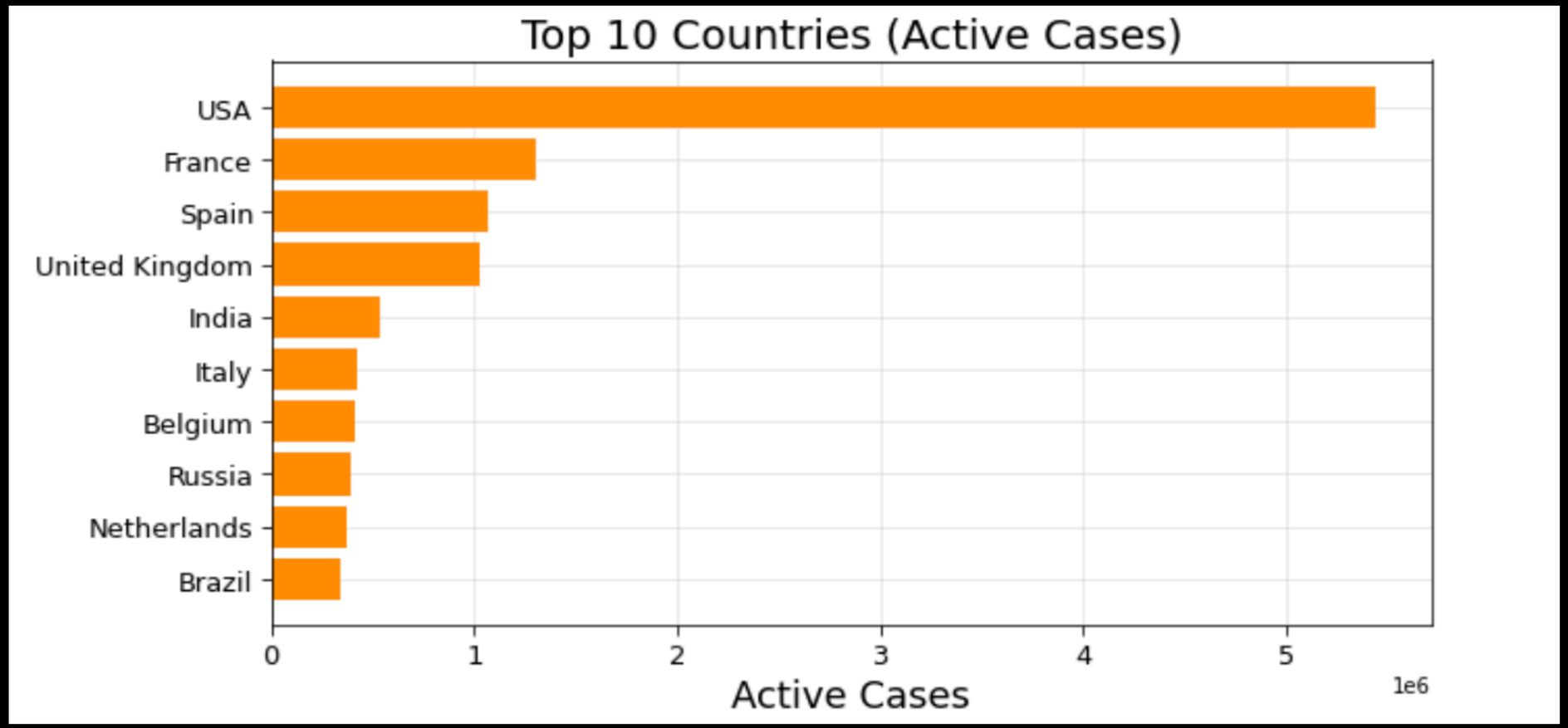
Top 10 COUNTRIES(CONFIRMED CASES)



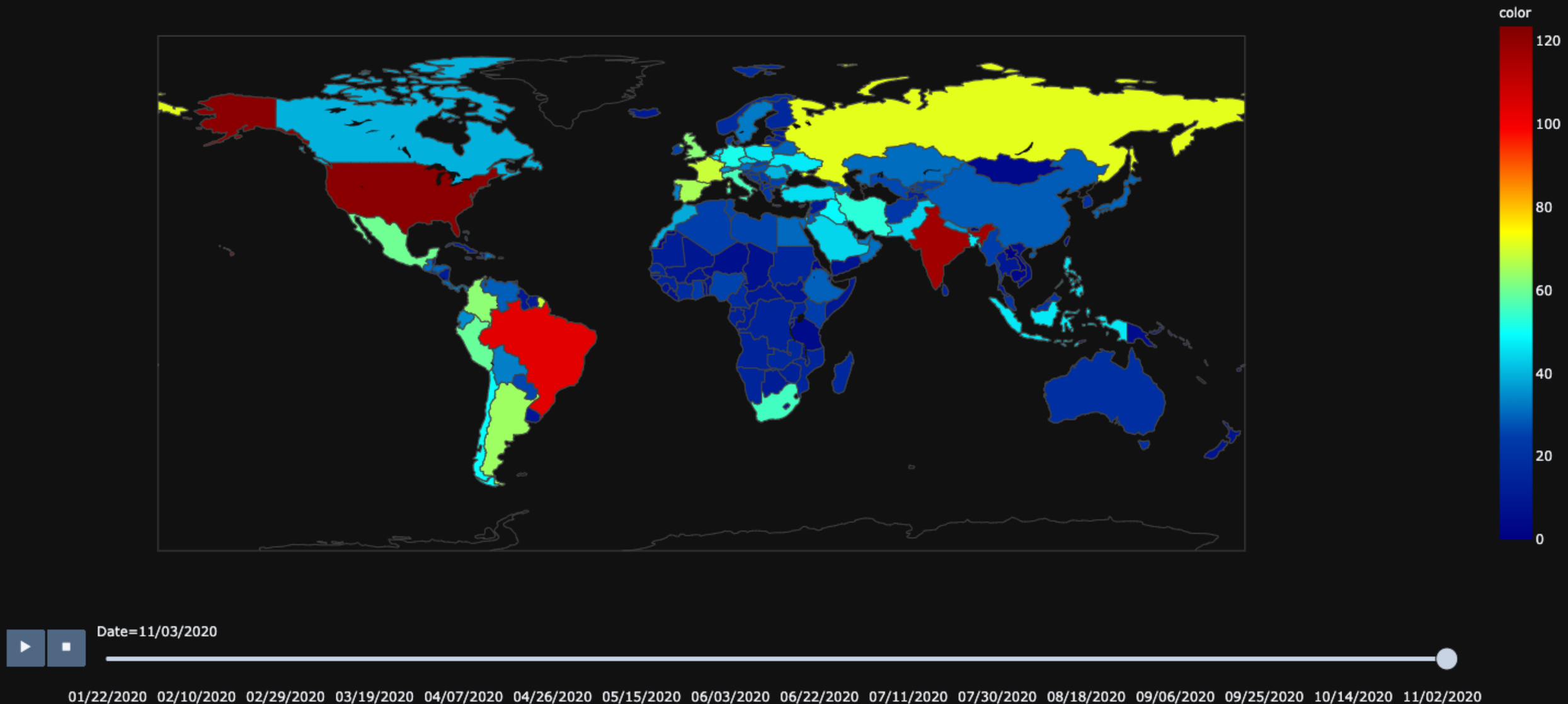
Top 10 Countries(Deaths Cases)



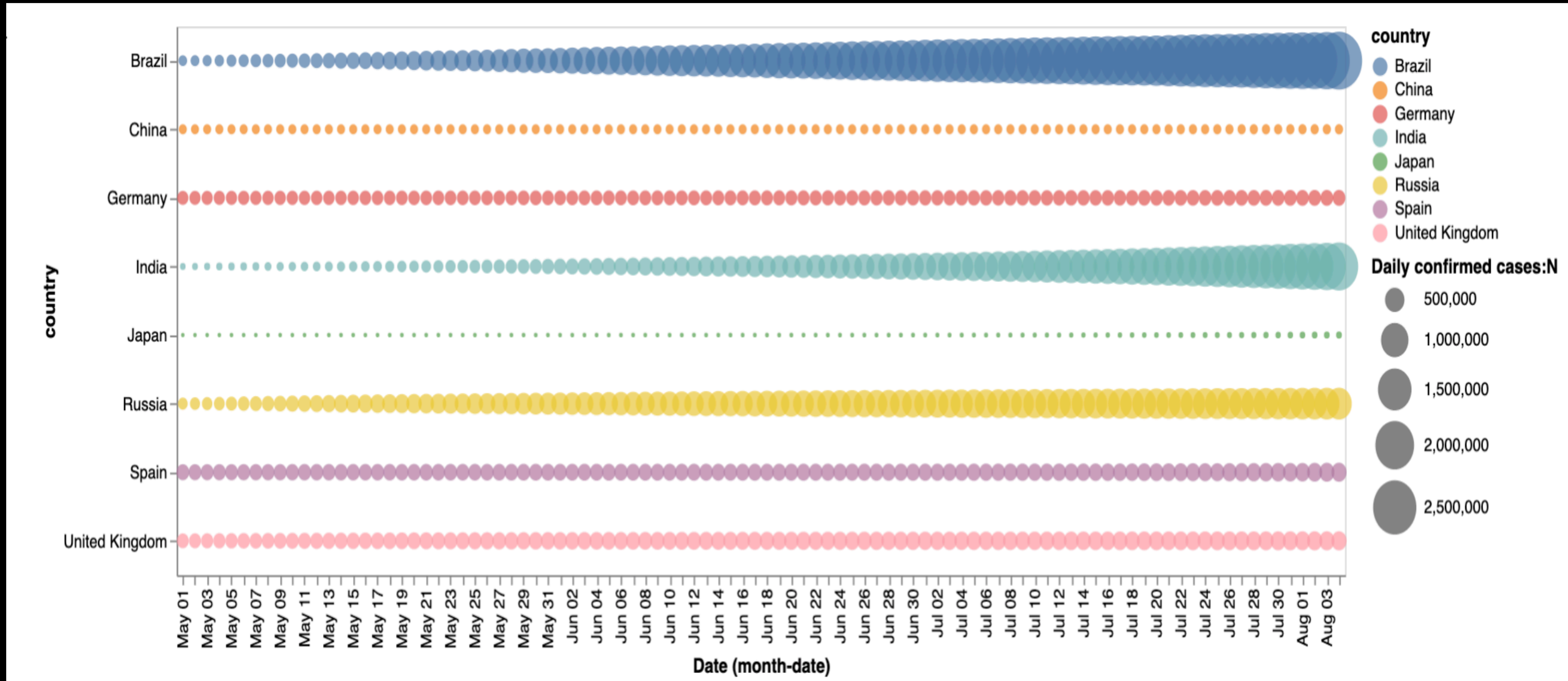
Top 10 Countries(Active Cases)

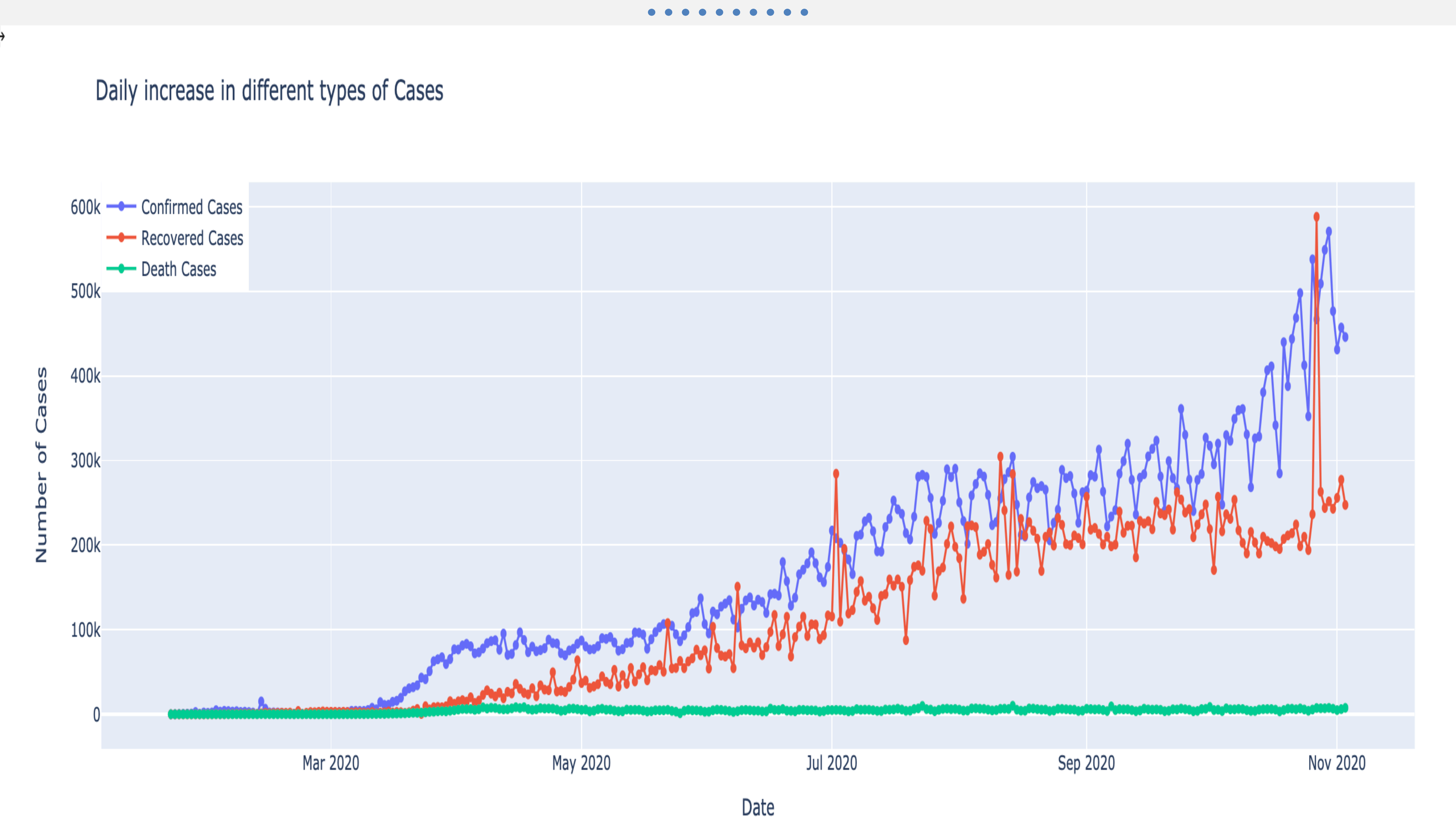


Covid-19:Progression of Spread

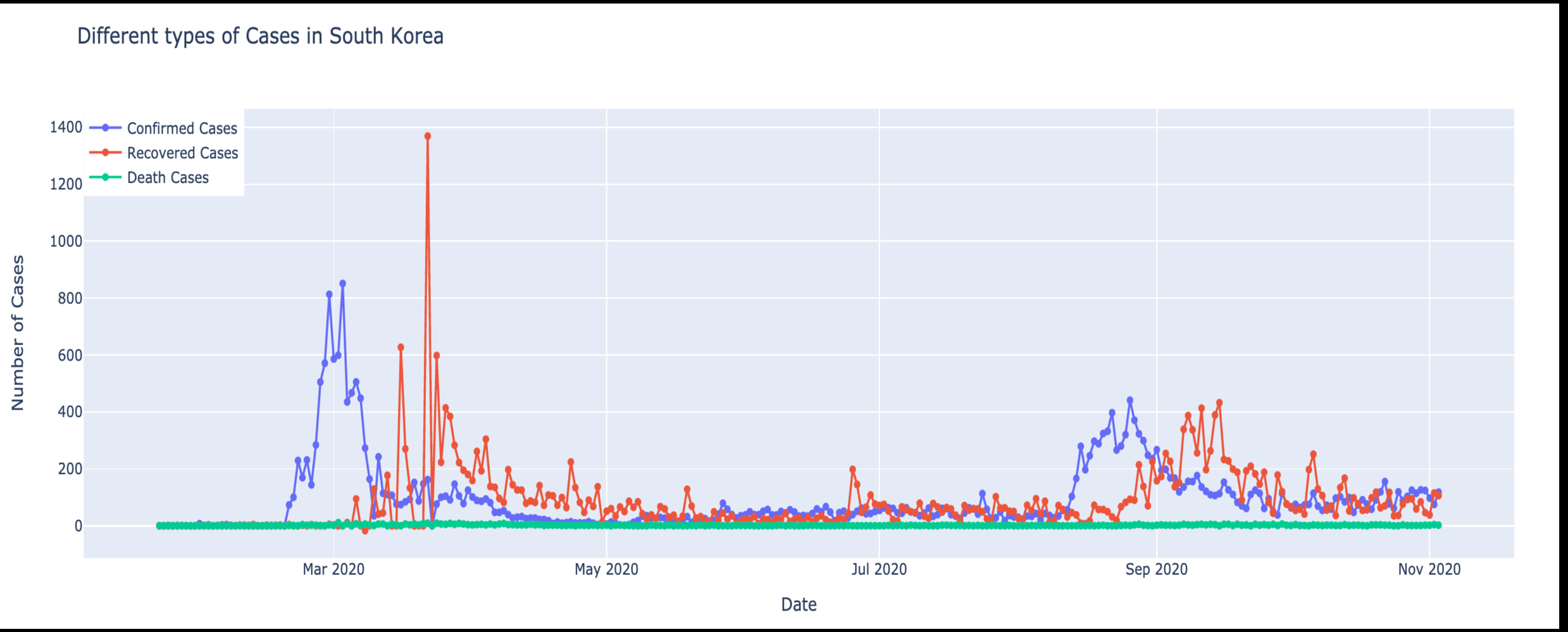


Comparison of spread of Covid-19 in different countries

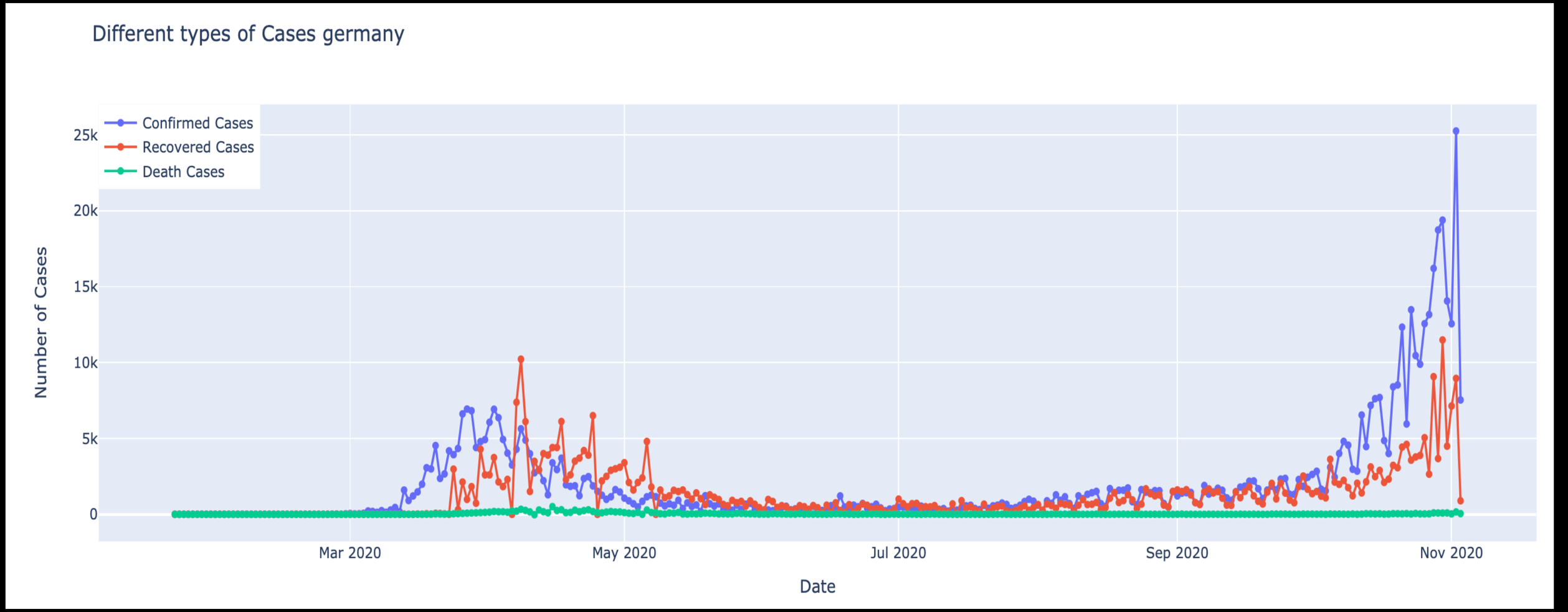



Daily increase in different types of Cases
Confirmed Cases
Recovered Cases
Death Cases
600k
500k
400k
300k
200k
100k
0
Number of Cases
Mar 2020
May 2020
Jul 2020
Sep 2020
Nov 2020
Date

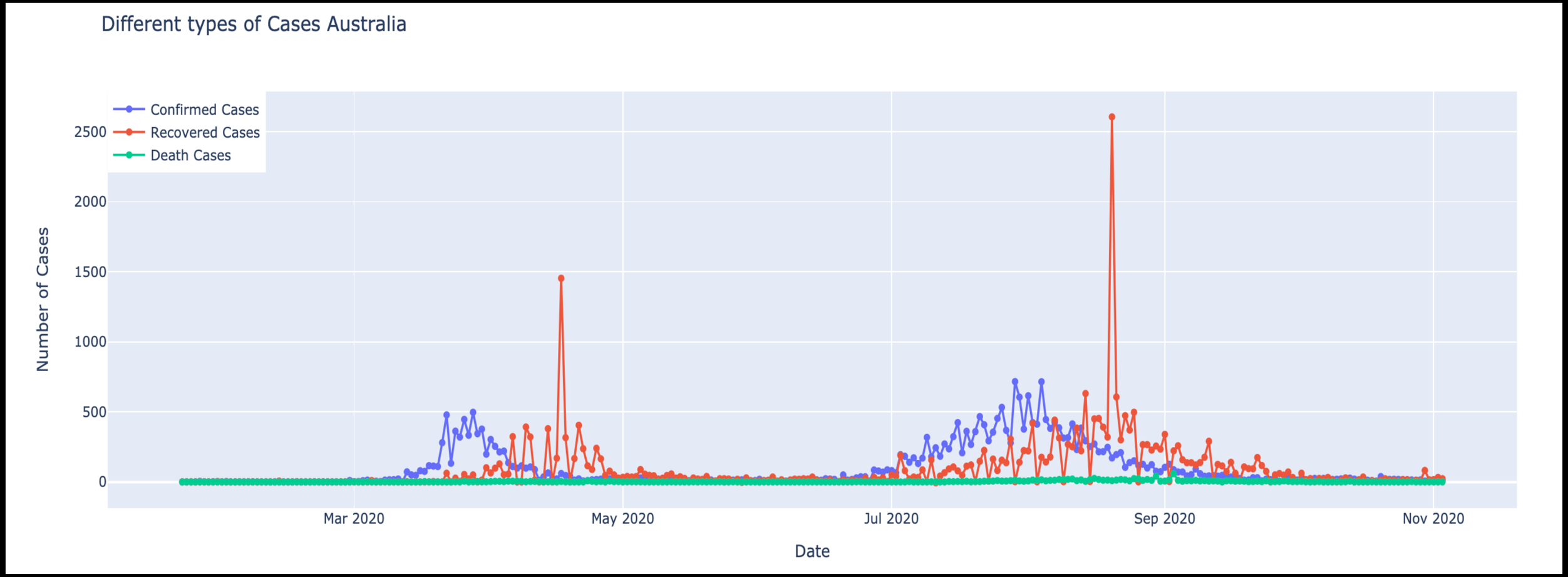
South Korea


Different types of Cases in South Korea
Confirmed Cases
Recovered Cases
Death Cases
Number of Cases
1400
1200
1000
800
600
400
200
0
Mar 2020
May 2020
Jul 2020
Sep 2020
Nov 2020
Date

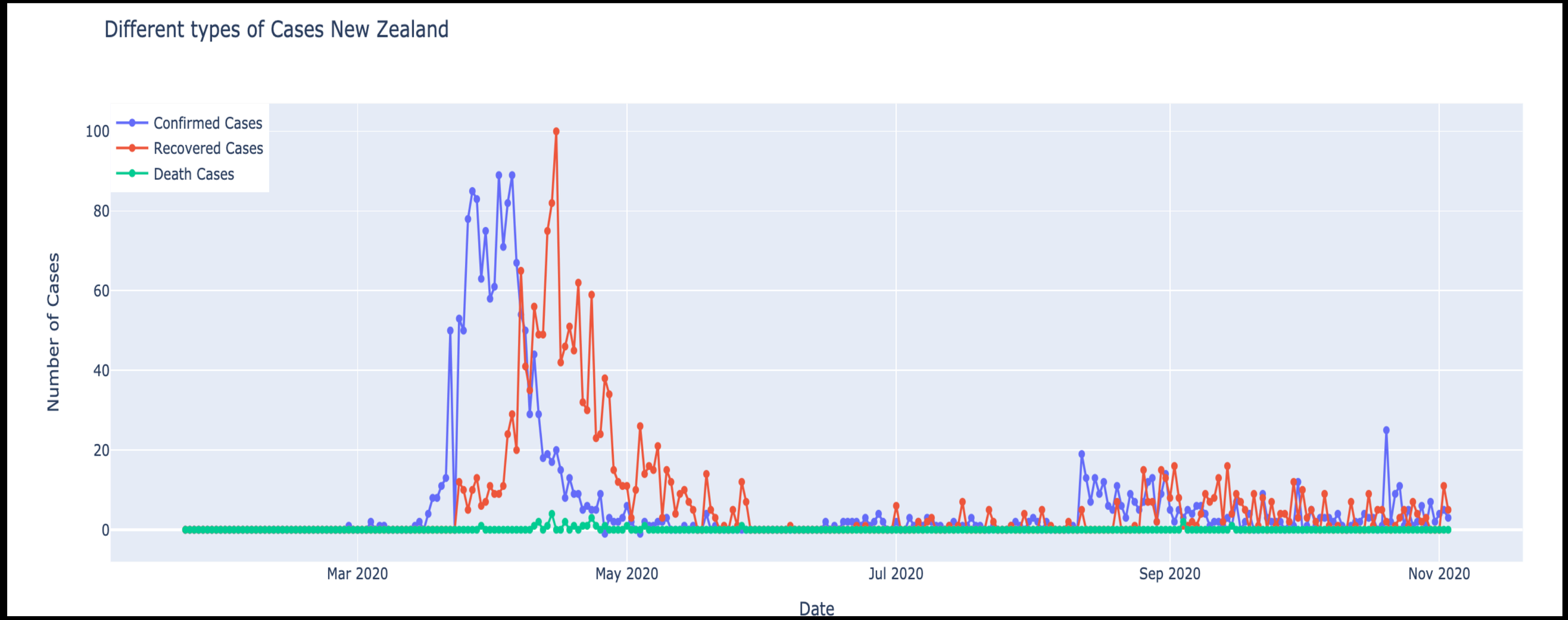
Germany

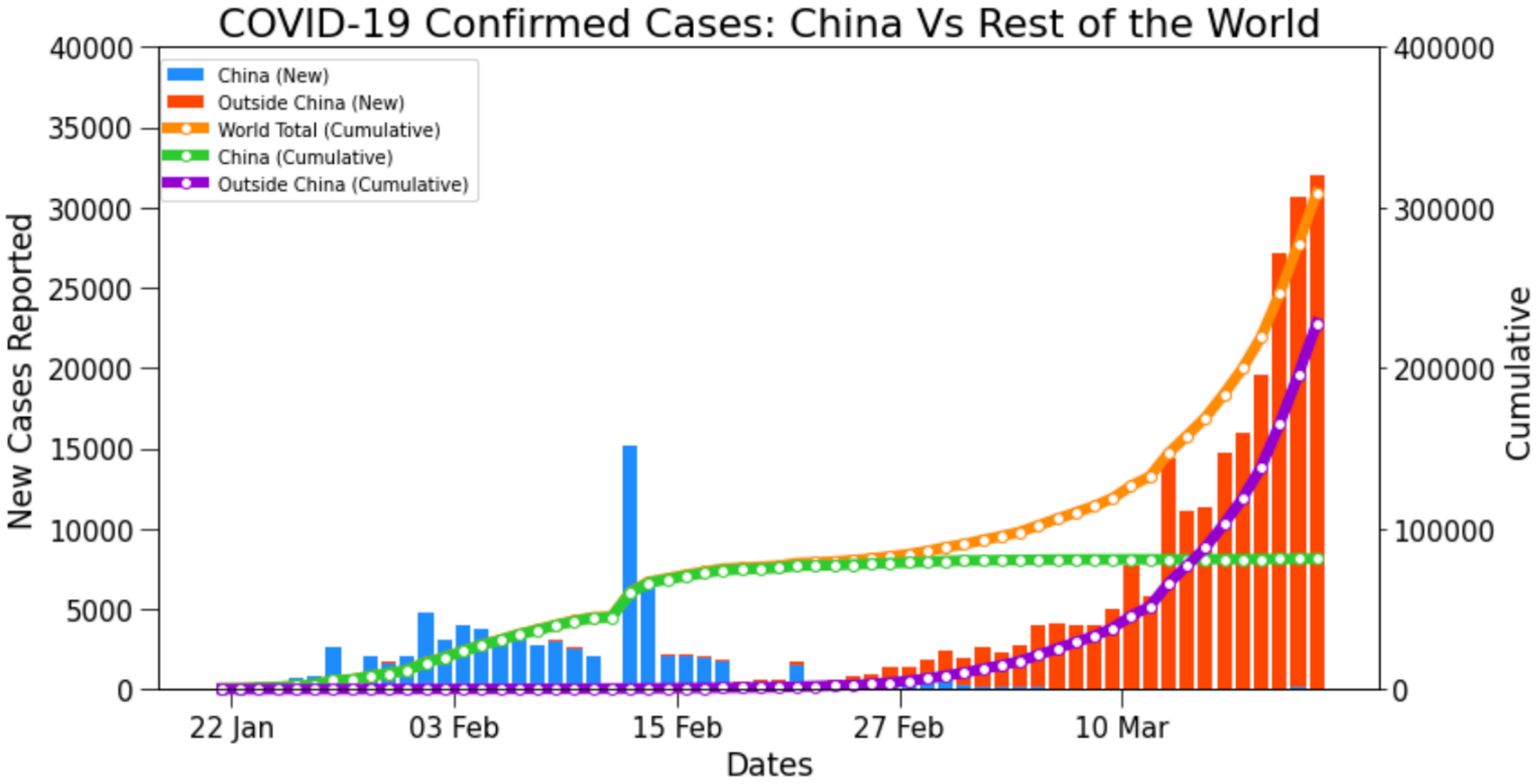

Different types of Cases germany
Confirmed Cases
Recovered Cases
Death Cases
25k
20k
15k
10k
5k
0
Number of Cases
Mar 2020
May 2020
Jul 2020
Sep 2020
Nov 2020
Date

Australia

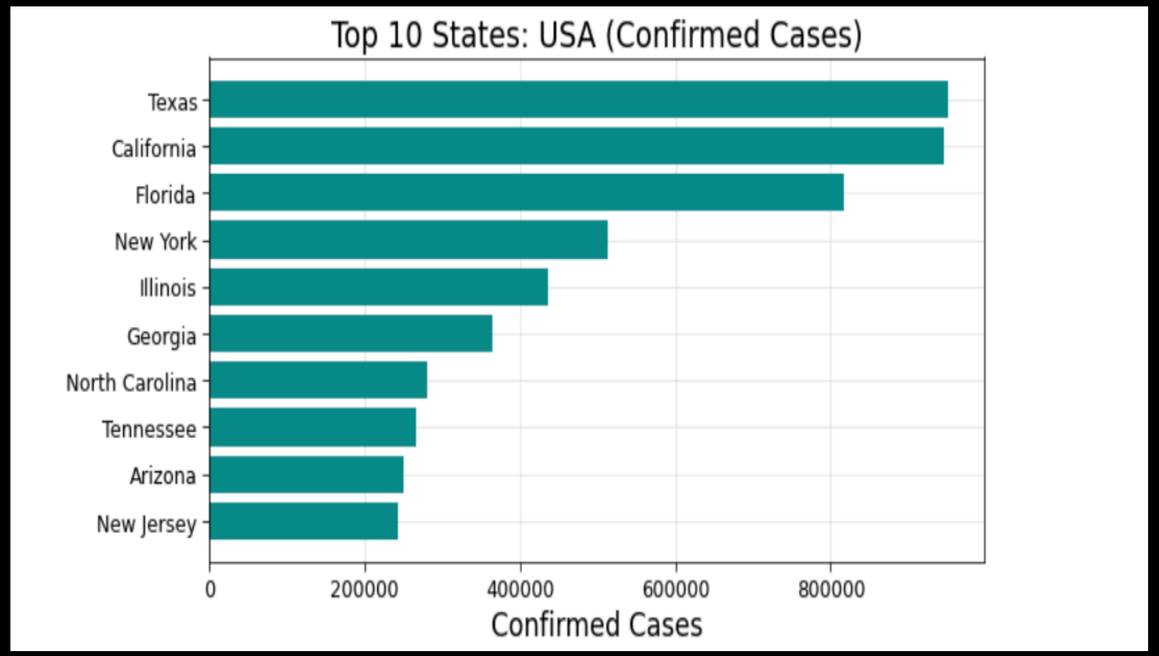

Different types of Cases Australia
Confirmed Cases
Recovered Cases
Death Cases
Number of Cases
2500
2000
1500
1000
500
0
Mar 2020
May 2020
Jul 2020
Sep 2020
Nov 2020
Date

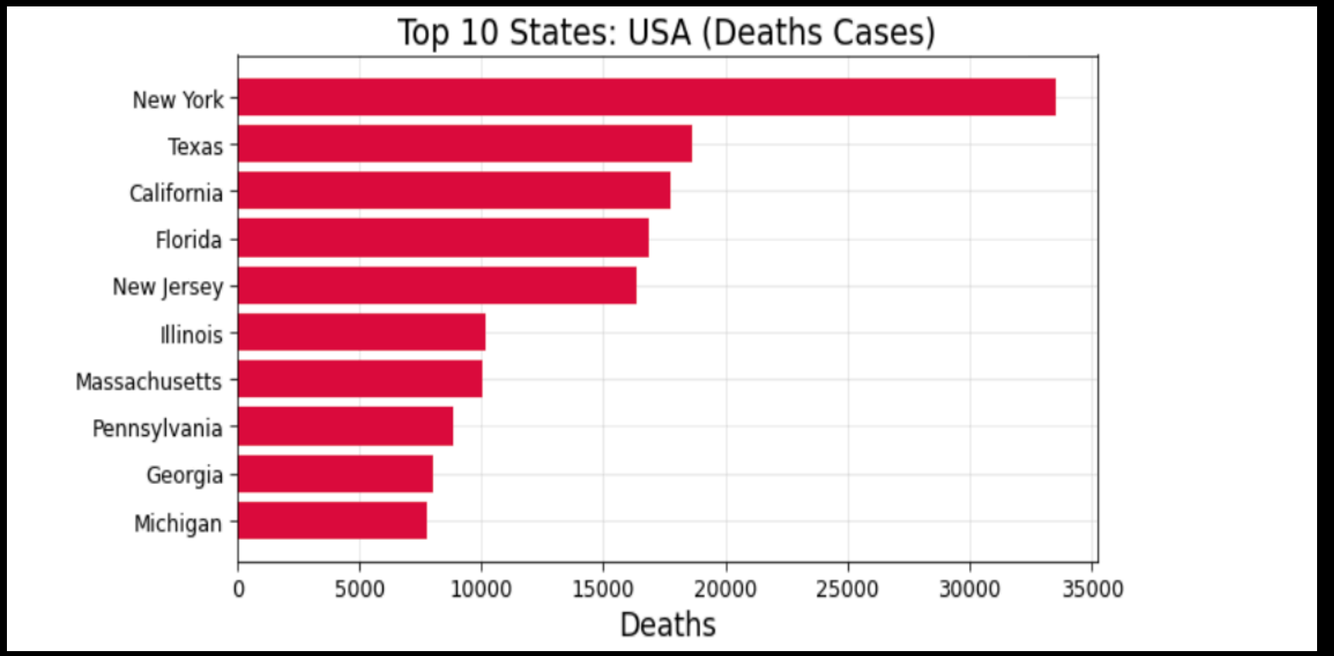
New Zealand


Different types of Cases New Zealand
Confirmed Cases
Recovered Cases
Death Cases
Number of Cases
100
80
60
40
20
0
Mar 2020
May 2020
Jul 2020
Sep 2020
Nov 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

USA


Top 10 States: USA (Confirmed Cases)
Texas
California
Florida
New York
Illinois
Georgia
North Carolina
Tennessee
Arizona
New Jersey
0
200000
400000
600000
800000
Confirmed Cases


Top 10 States: USA (Deaths Cases)
New York
Texas
California
Florida
New Jersey
Illinois
Massachusetts
Pennsylvania
Georgia
Michigan
0
5000
10000
15000
20000
25000
30000
35000
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

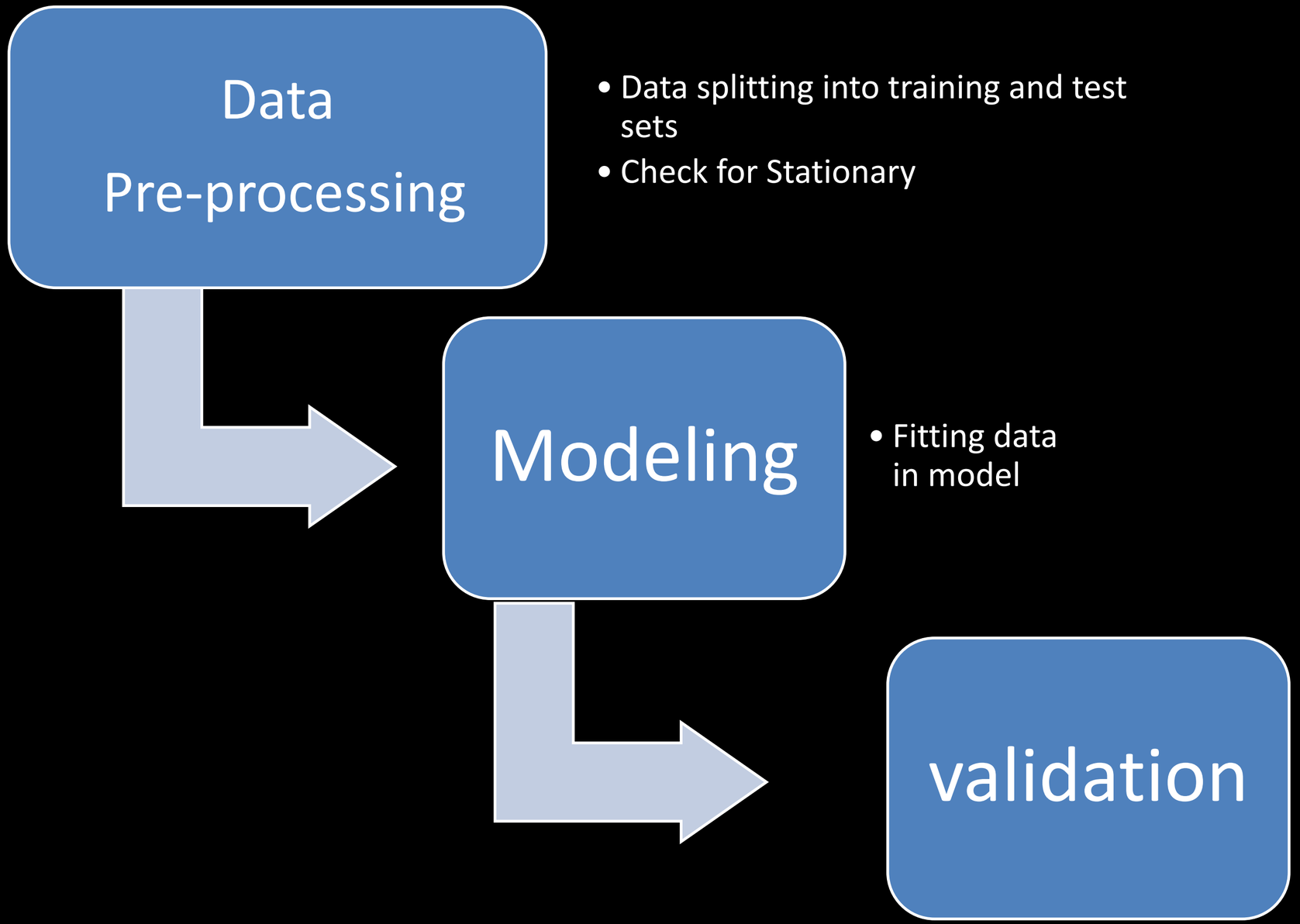
  2.ARIMA

  3.SARIMA

  4.PROPHET

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

Modeling steps



	Model Name	Root Mean Squared Error
2	SARIMA Model	5.949517e+04
0	Holt's Winter Model	6.925728e+04
1	ARIMA Model	5.699028e+05
3	Facebook's Prophet Model	4.411156e+07

Models Comparison

- RMSE of SARIMA Model has good accuracy with RMSE 5.9495 and AIC score 6017 so I will be using SARIMA model to forecast covid19 cases.

SOME DETAILS ON BEST MODEL(SARIMA)

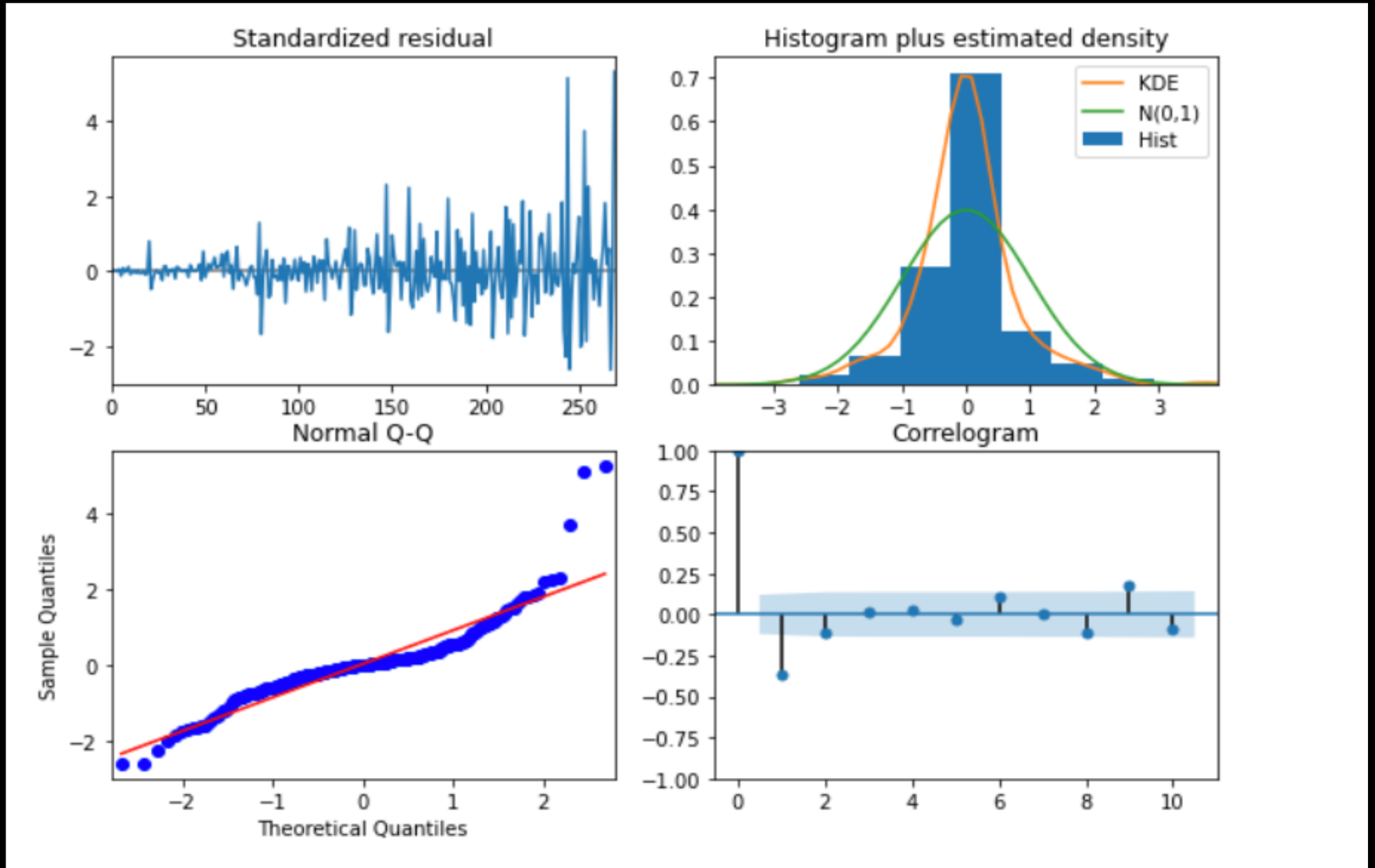
SARIMAX Results

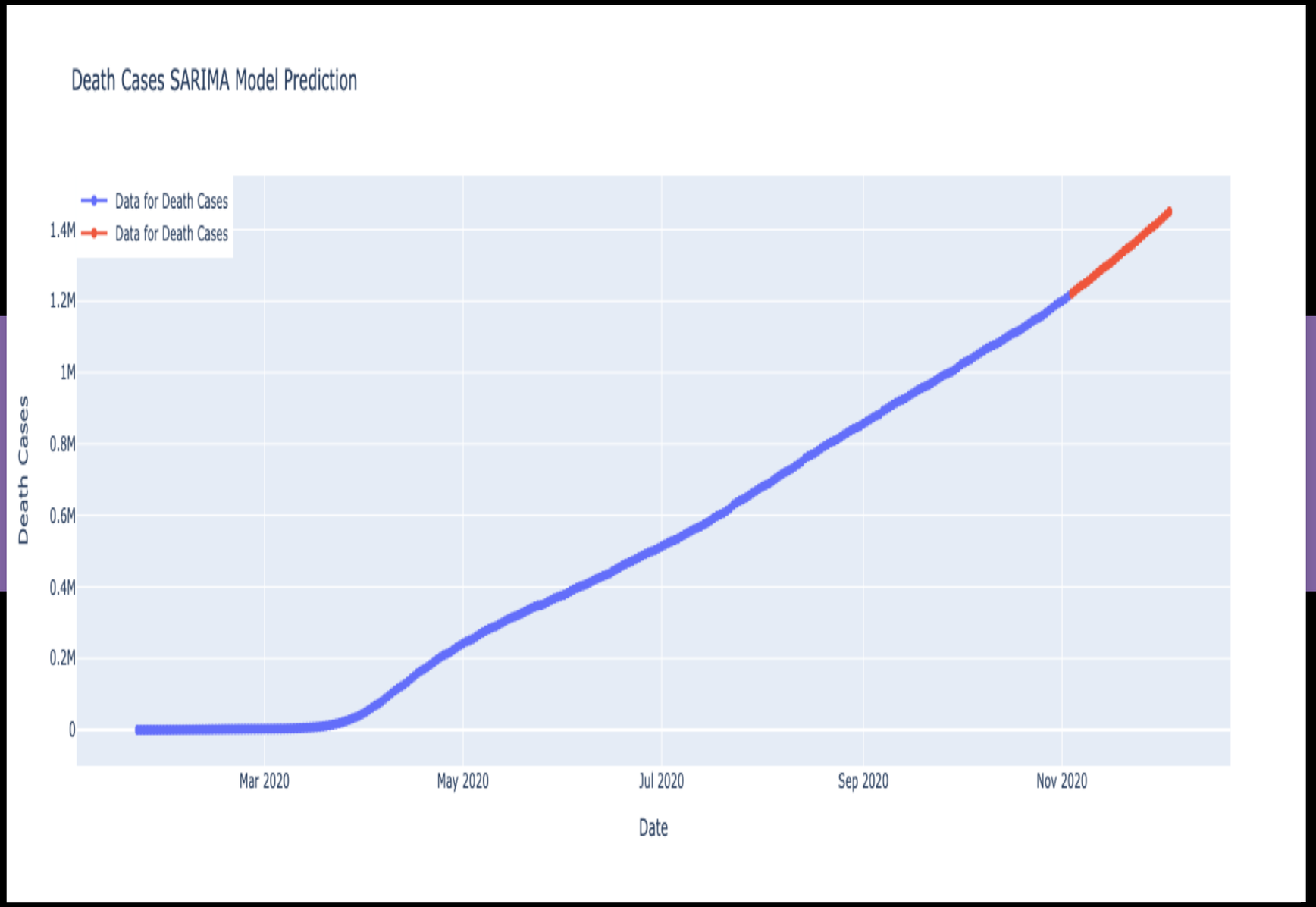
Dep. Variable:	y	**No. Observations:**	272
Model:	SARIMAX(0, 2, 0)x(2, 0, [1, 2], 7)	**Log Likelihood**	-3003.881
Date:	Thu, 05 Nov 2020	**AIC**	6017.761
Time:	07:19:00	**BIC**	6035.753
Sample:	0	**HQIC**	6024.986
	- 272		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.S.L7	1.8039	0.293	6.157	0.000	1.230	2.378
ar.S.L14	-0.8163	0.295	-2.769	0.006	-1.394	-0.238
ma.S.L7	-1.5551	0.278	-5.590	0.000	-2.100	-1.010
ma.S.L14	0.6804	0.176	3.871	0.000	0.336	1.025
sigma2	3.177e+08	2.04e-09	1.56e+17	0.000	3.18e+08	3.18e+08

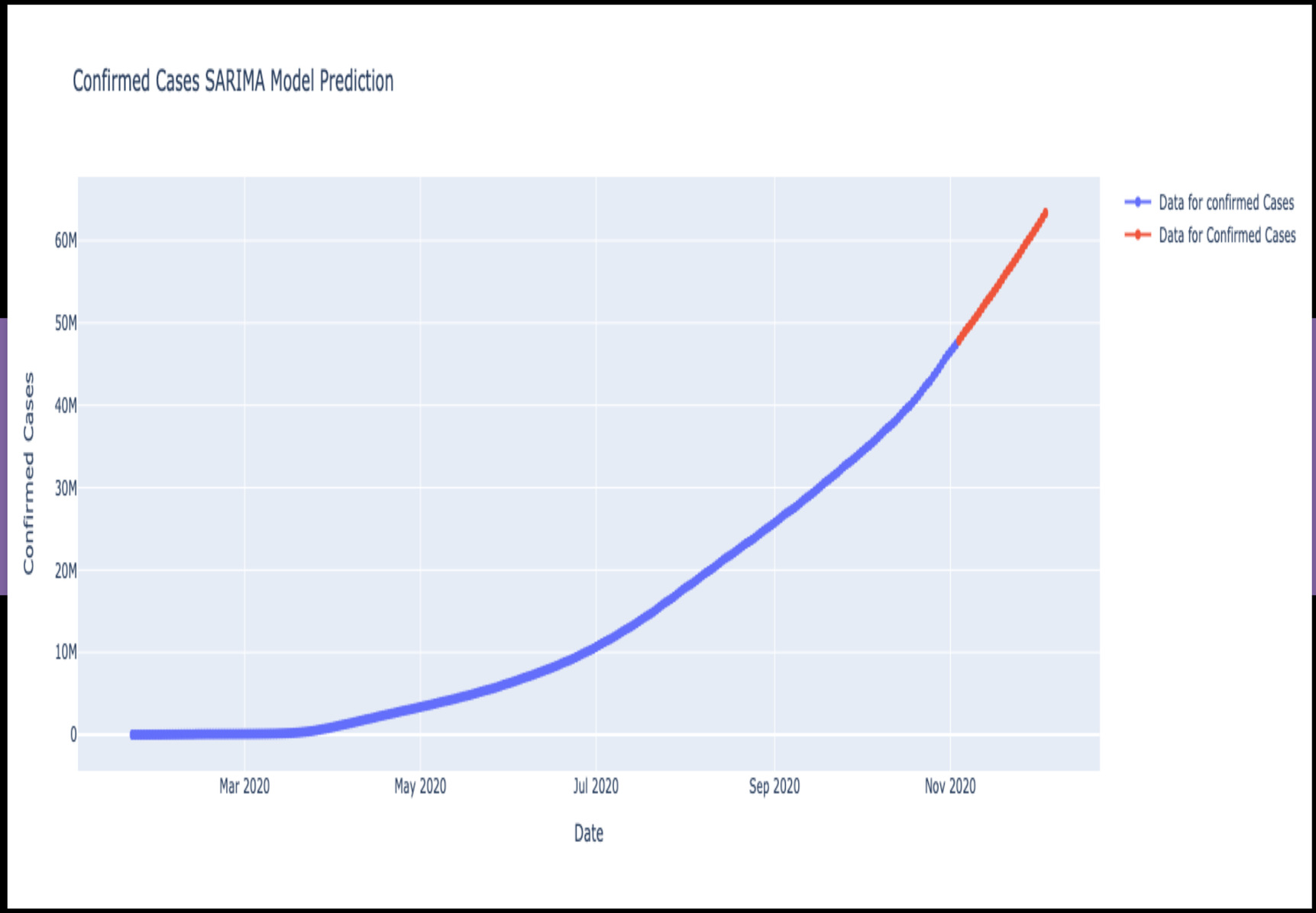
Ljung-Box (Q):	93.67	**Jarque-Bera (JB):**	1028.70
Prob(Q):	0.00	**Prob(JB):**	0.00
Heteroskedasticity (H):	16.94	**Skew:**	1.57
Prob(H) (two-sided):	0.00	**Kurtosis:**	12.03

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 5.9495 and AIC score 6017 Model has forecasted increase of death cases to 14,500,540 and confirmed cases to 63,34,800 by 2020-12-04 worldwide.