Data-Driven Exploration of Covid-19 Cases and Deaths Trending

Francis Perez, Xinrui Zhang, Raiana Zaman, Jamison Valentine, Jason Manning

University of North Carolina at Greensboro

Greensboro, NC

Abstract—The goal of this project is to develop an analytical framework that can be used to study patterns of COVID-19's effect and spread throughout the United States.

I. Introduction

This project uses data to generate a detailed analysis of daily and weekly trends of Covid-19 cases and deaths.

II. Load Data and Intuition

A. Get Input Data type from Datasets

First, download datasets from the providing website, and load datasets using pandas package command. Get data types from each datasets loaded.

Covid_county_population_usafacts.csv		
This file contains a list of counties in the U.S.A with their related populations.		
Variable	**Data Type**	**Description**
county FIPS	int64	The FIPS id of the county. A Unique code for each county of the U.S.A
County Name	object/string	The name of the county.
State	object/string	The name of the state.
population	int64	The population of the county.

covid_confirmed_usafacts (known cases).csv		
This file contains a list of counties in the U.S.A with their related confirmed cases of Covid-19. The file contains ever growing columns of dates with a running total of confirmed cases.		
Variable	**Data Type**	**Description**
county FIPS	int64	The FIPS id of the county. A Unique code for each county in the U.S.A
County Name	object/string	The name of the county.
State	object/string	The name of the state.
state FIPS	int64	The FIPS id of the state. A Unique code for each state in the U.S.A
1/22/2020 1/23/2020 … 9/7/2020	int64	The running total of confirmed cases starting on 1/22/2020.

covid_deaths_usafacts.csv		
This file contains a list of counties in the U.S.A with their related confirmed cases of Covid-19. The file contains ever growing columns of dates with a running total of deaths.		
Variable	**Data Type**	**Description**
county FIPS	int64	The FIPS id of the county. A Unique code for each county in the U.S.A
County Name	object/string	The name of the county.
State	object/string	The name of the state.
state FIPS	int64	The FIPS id of the state. A Unique code for each state in the U.S.A
1/22/2020 1/23/2020 … 9/7/2020	int64	The running total of deaths starting on 1/22/2020.

B. Merge datasets

Merge data from each dataset and form an enrichment super dataset and make intuitions on how to utilize the enrichment data.

III. Data Utilization

Utilize statistical modeling to develop formal hypothesis tests for the intuitions made in stage 1.

A. Compare Statistics

Compare the weekly statistics (mean, median, mode) for the number of new cases and deaths across the US. You are calculating the mean (rounded to integer value) number of new cases and per week and then calculating (mean, median, mode) for all weeks taken together.

```
Confirmed Cases Mean: 29,450
Confirmed Cases Median: 28,011
Confirmed Cases Mode: 1
Deaths Mean: 794
Deaths Median: 741
Deaths Mode: 0
```

The mean and median number of new cases were fairly close, with the median being less than the mean, indicating a slight right-skew in the data. Such was also the case with the number of new reported deaths. The respective modes the most frequently occuring observations were 1 new confirmed case and 0 deaths across the US.

B. Compare between Countries

list of countries that we are going to compare

1. Brazil in South America
2. Japan in Asia
3. Mexico in North America
4. Nigeria in Africa
5. Russia in Europe

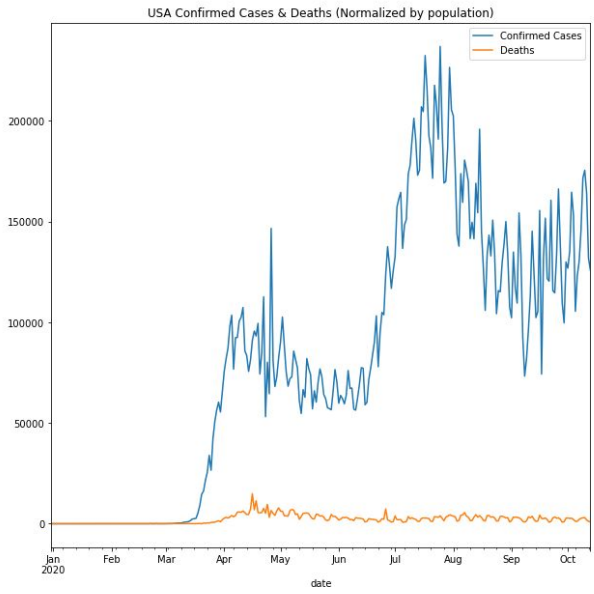
These countries were chosen as a way to represent each of the world land area & similar populations

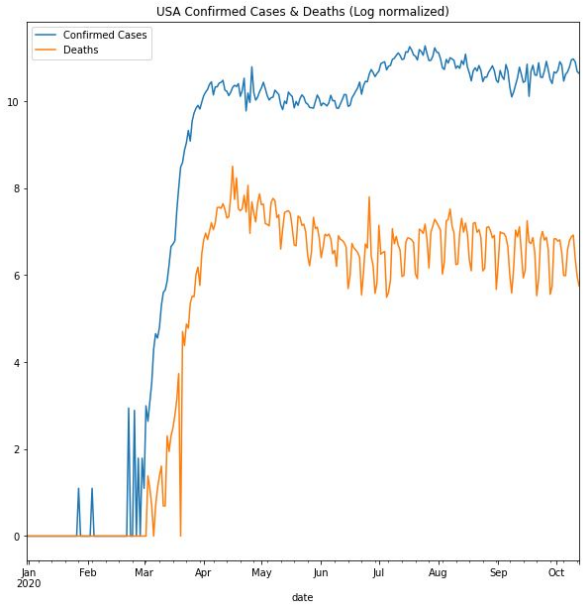
	Confirmed Cases Mean	Confirmed Cases Median	Confirmed Cases Mode	Deaths Mean	Deaths Cases Median	Deaths Cases Mode
Brazil	82,495	64,338	0	2,431	2,954	0

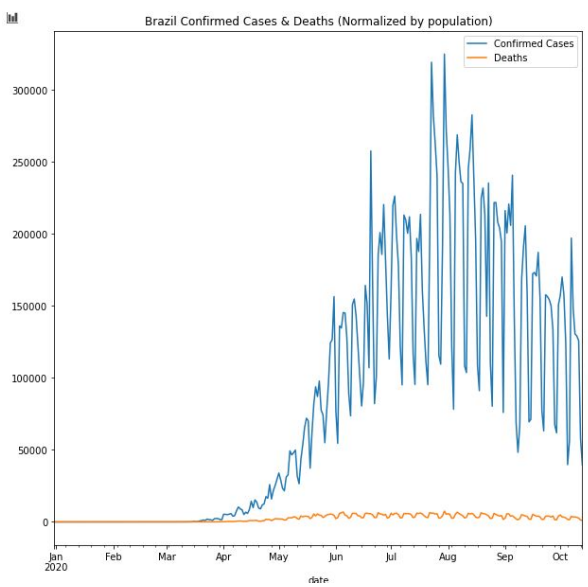
Japan	2,463	1,152	0	44	24	0
Mexico	22,103	22,364	0	2,235	2,120	0
Nigeria	1,011	774	0	18	10	0
Russia	32,173	36,198	0	545	698	0

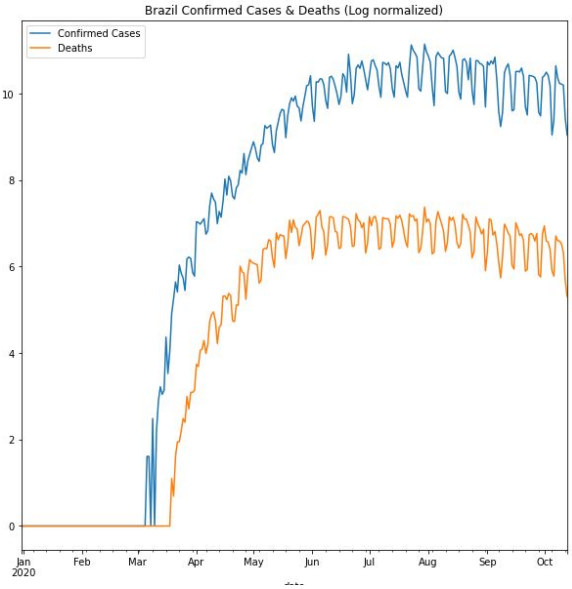
C. Data Normalization and plots

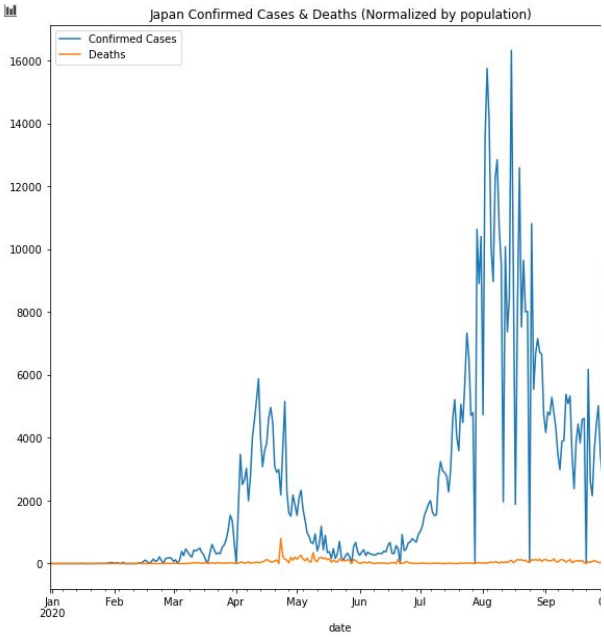
After normalization by population, we plot the normalized and log-normalized confirmed cases and deaths figures based on each countries' data.

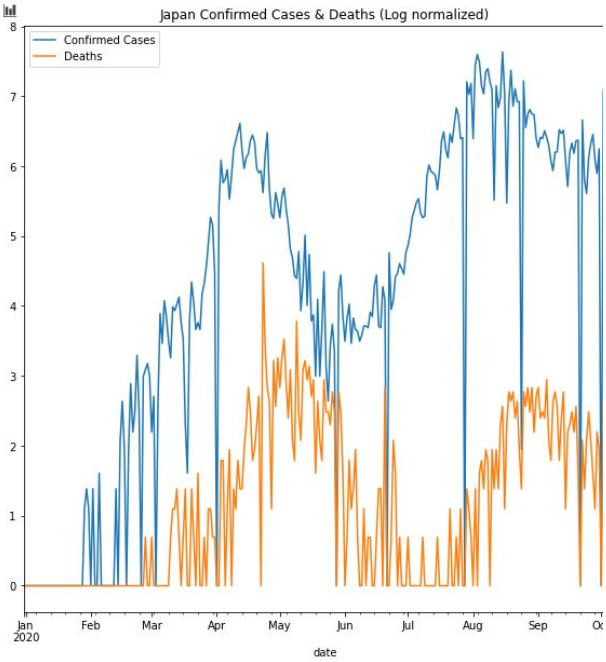


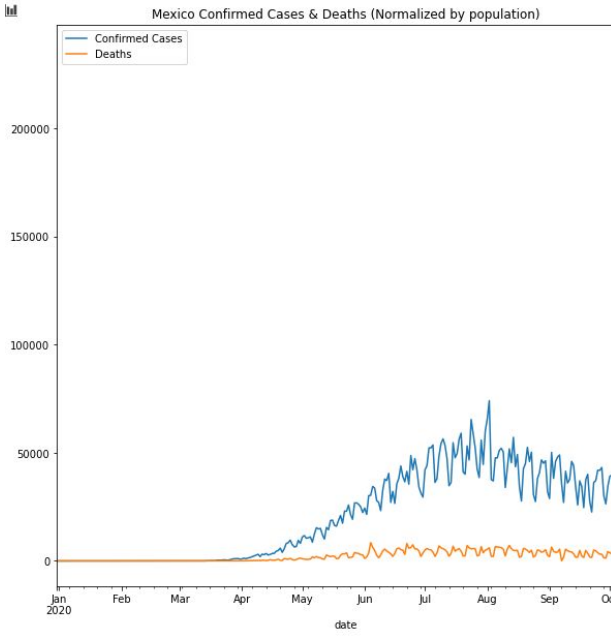


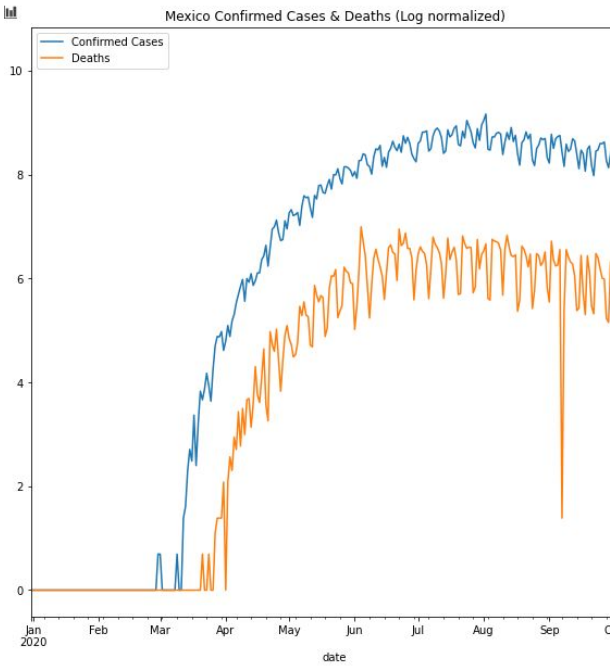

Brazil Confirmed Cases & Deaths (Normalized by population)
Confirmed Cases
Deaths
300000
250000
200000
150000
100000
50000
0
Jan
2020
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
date

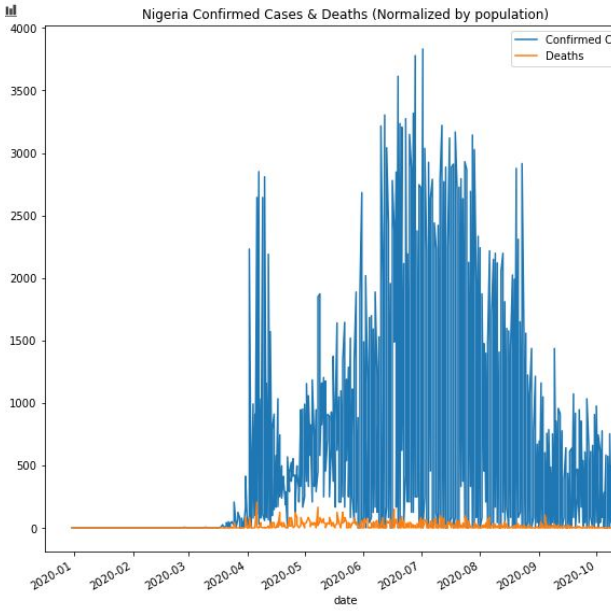

Brazil Confirmed Cases & Deaths (Log normalized)
Confirmed Cases
Deaths
10
8
6
4
2
0
Jan
2020
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct

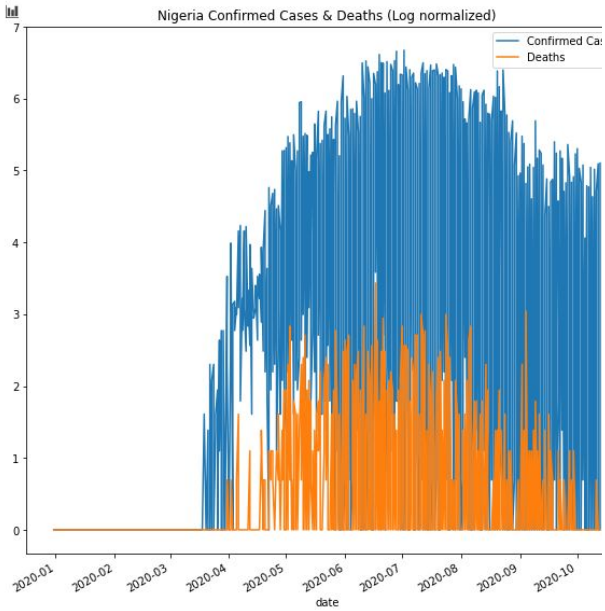

Japan Confirmed Cases & Deaths (Normalized by population)
Confirmed Cases
Deaths
16000
14000
12000
10000
8000
6000
4000
2000
0
Jan
2020
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
date

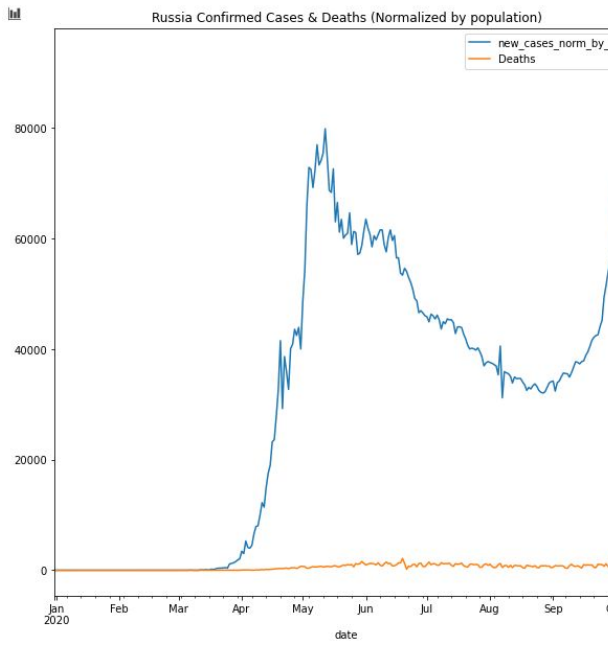

Japan Confirmed Cases & Deaths (Log normalized)
Confirmed Cases
Deaths
8
7
6
5
4
3
2
1
0
Jan
2020
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
date

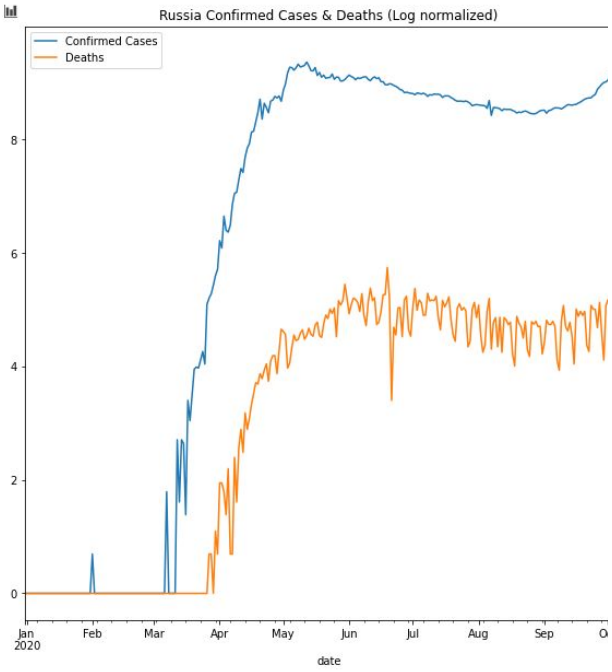

Mexico Confirmed Cases & Deaths (Normalized by population)
Confirmed Cases
Deaths
200000
150000
100000
50000
0
Jan
2020
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
date


Mexico Confirmed Cases & Deaths (Log normalized)
Confirmed Cases
Deaths
10
8
6
4
2
0
Jan
2020
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
date


Nigeria Confirmed Cases & Deaths (Normalized by population)
Deaths
4000
3500
3000
2500
2000
1500
1000
500
0
2020-01
2020-02
2020-03
2020-04
2020-05
2020-06
2020-07
2020-08
2020-09
2020-10
date


Nigeria Confirmed Cases & Deaths (Log normalized)
Deaths
7
6
5
4
3
2
1
0
2020-01
2020-02
2020-03
2020-04
2020-05
2020-06
2020-07
2020-08
2020-09
2020-10
date





D. Identify peak week of the cases and deaths in US and other countries

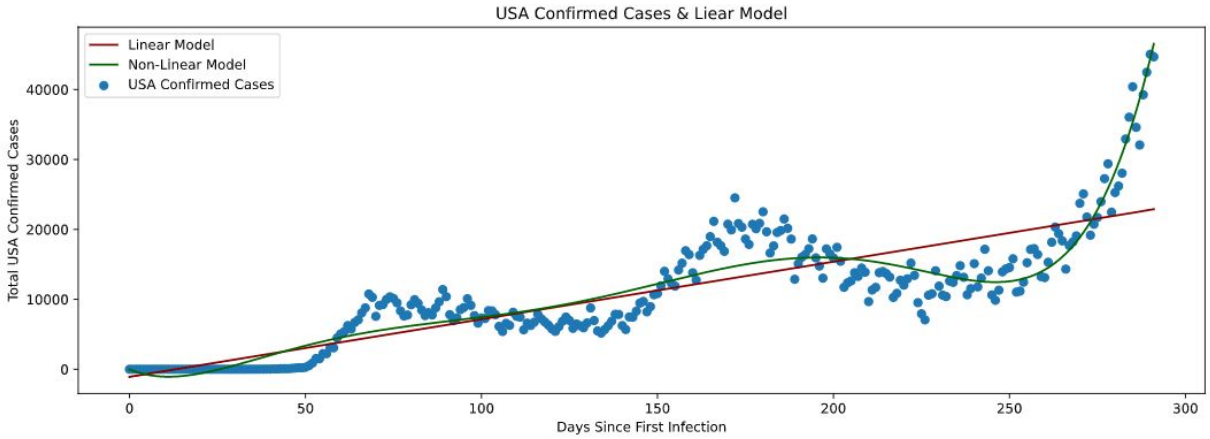
Listed below are the statistics for the peak weeks for each of the following countries: USA, Brazil, Japan, Mexico, Nigeria, Russian. Based on the data below, Russian appears to have the worst week overall for new cases and deaths. Nigeria appears to have fared better on their worst week than the other countries.

```
USA     - Peak Week - New Cases:30  | New Deaths:  16
Brazil  - Peak Week - New Cases:30  | New Deaths:  30
Japan   - Peak Week - New Cases:32  | New Deaths:  17
Mexico  - Peak Week - New Cases:41  | New Deaths:  26
Nigeria - Peak Week - New Cases:26  | New Deaths:  25
Russia  - Peak Week - New Cases:41  | New Deaths:  41
```

IV. Regression Model

Apply different regression models on data to predict and analyze daily trends of Covid-19 cases and deaths in the US. And then compare with other countries.

A. Linear and Non-linear models with prediction



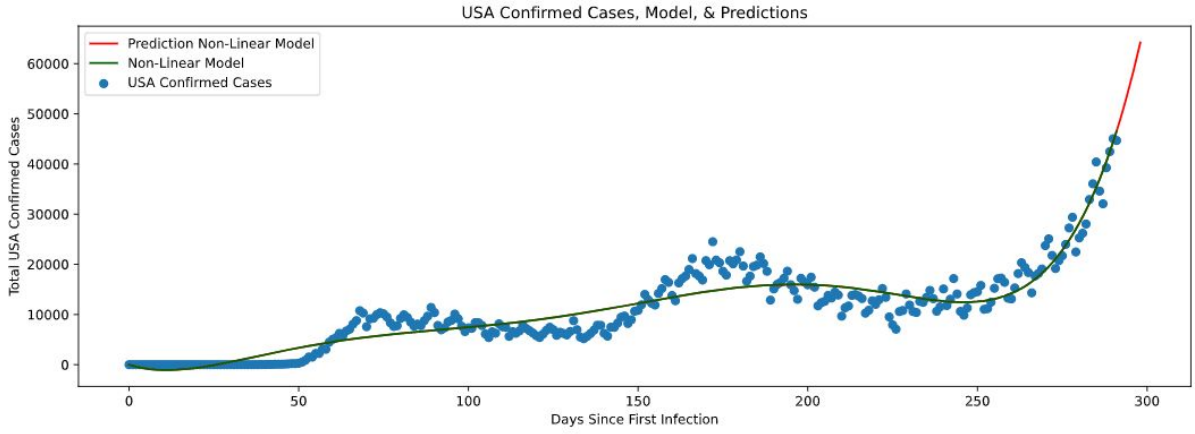
*Normalize By Population Per 100,000,000

Confirmed Cases - RMSE - Linear Model: 4828.14 | Non-Linear Model: 2799.03

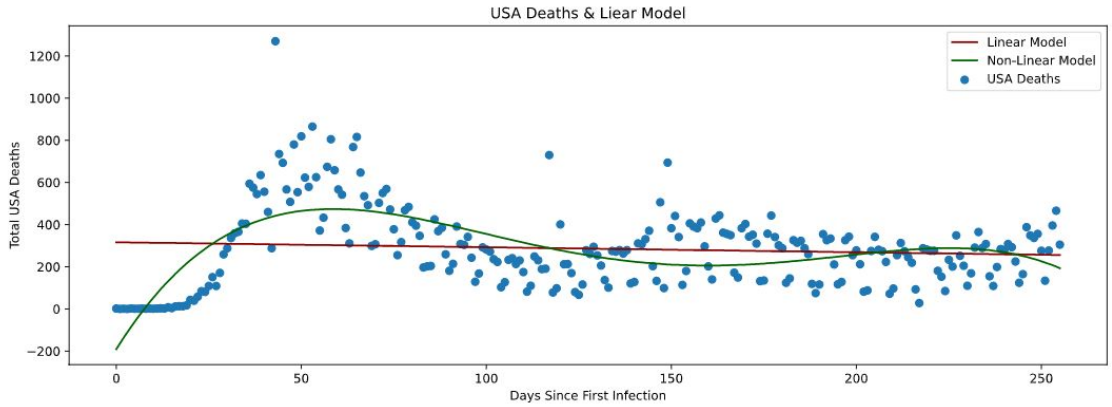
We used a NonLinear Model for the predictions on a lower RMSE value (above).

******Number of New Confirmed Cases 7-Days in the future.*******

days_since_infection_start	daily_count
292	48,740
293	51,043
294	53,452
295	55,970
296	58,599
297	61,344
298	64,209



*Normalize By Population Per 100,000,000



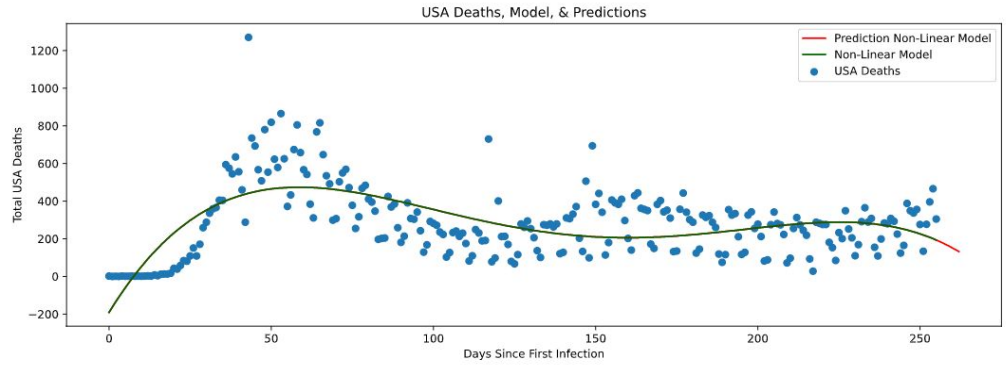
*Normalize By Population Per 100,000,000

Deaths - RMSE - Linear Model: 190.42 | Non-Linear Model: 150.32

We used a NonLinear Model for the predictions on a lower RMSE value (above).

******Number of New Deaths 7-Days in the future.*******

days_since_infection_start	daily_count
256	186
257	178
258	169
259	161
260	151
261	142
262	132



*Normalize By Population Per 100,000,000

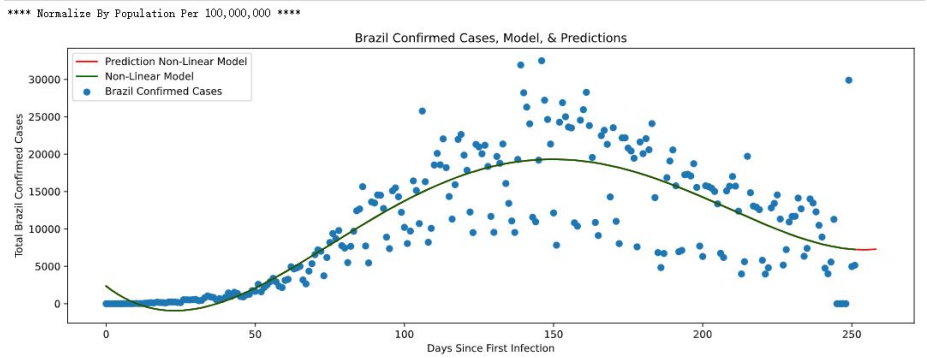
B. Compare with other countries

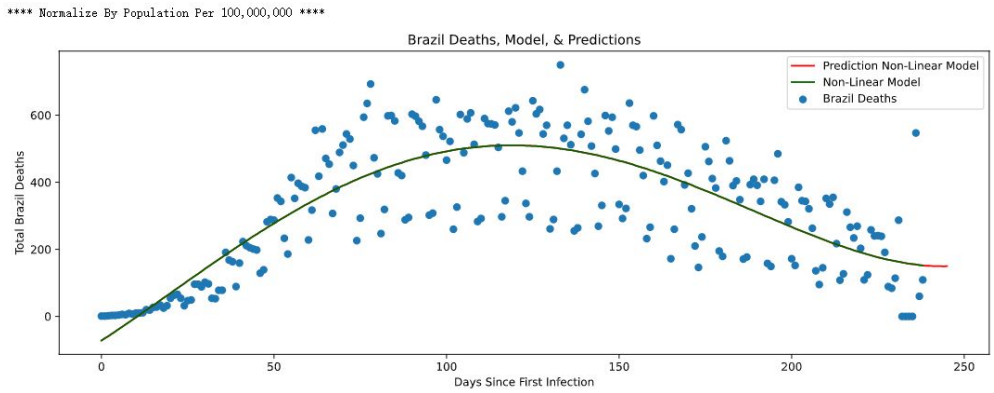
USA

- Confirmed Cases: The USA is predicted to see a sharp increase in the number of new Confirmed cases in the week of 11-13-2020 to 11-20-2020.

- Deaths: The USA is predicted to see a slight decrease in the number of deaths in the week of 11-13-2020 to 11-20-2020.

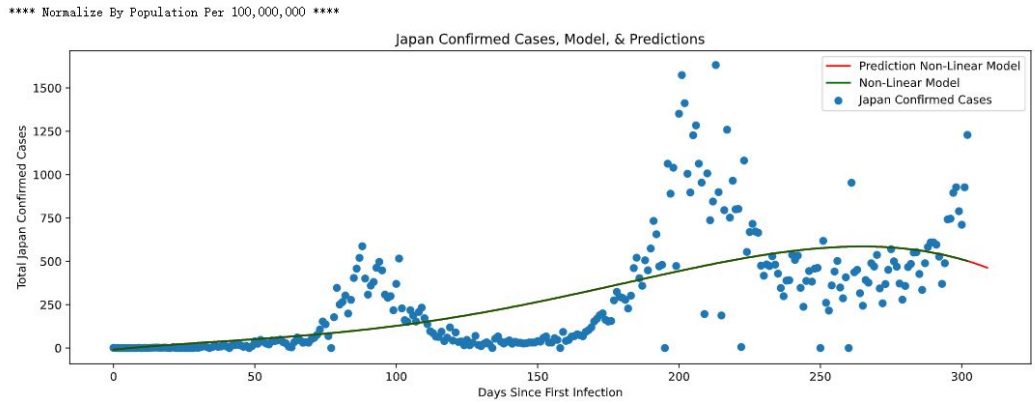
Brazil



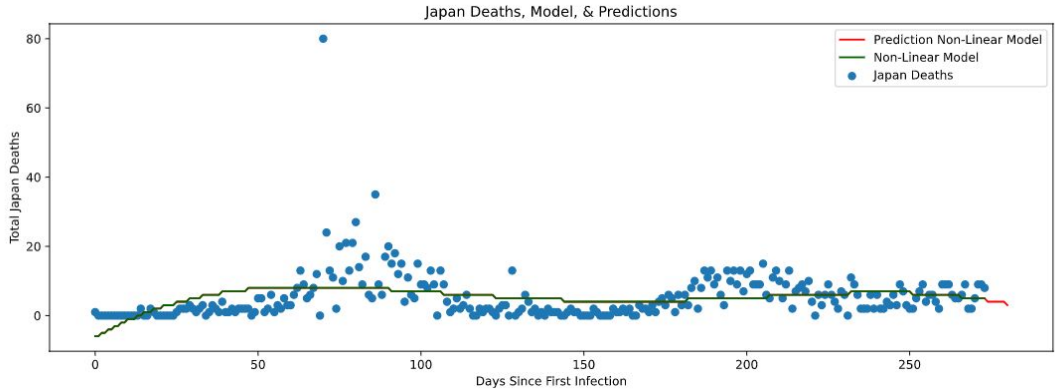


- Confirmed Cases: Brazil is predicted to see very little change, holding at around 7200 per 100,000,000 population, of new Confirmed cases in the week of 11-13-2020 to 11-20-2020.
- Deaths: Brazil is predicted to see very little change, holding at around 150 per 100,000,000 population, of new deaths in the week of 11-13-2020 to 11-20-2020.

Japan



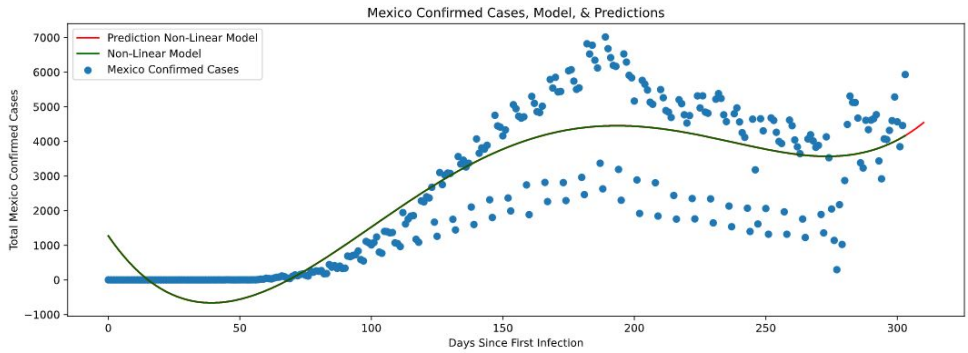
**** Normalize By Population Per 100,000,000 ****



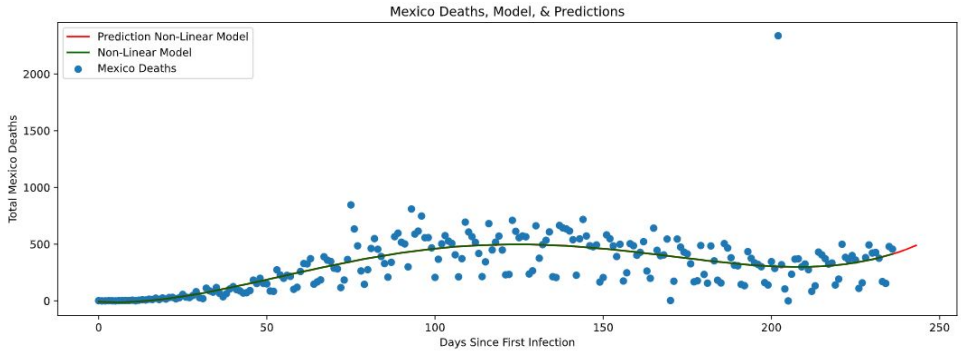
- Confirmed Cases: Japan is predicted to see very little change, holding at around 300 per 100,000,000 population, of new Confirmed cases in the week of 11-13-2020 to 11-20-2020.
- Deaths: Japan is predicted to see very little change, hold at around 4 per 100,000,000 population, of new deaths in the week of 11-13-2020 to 11-20-2020.

Mexico

**** Normalize By Population Per 100,000,000 ****



**** Normalize By Population Per 100,000,000 ****

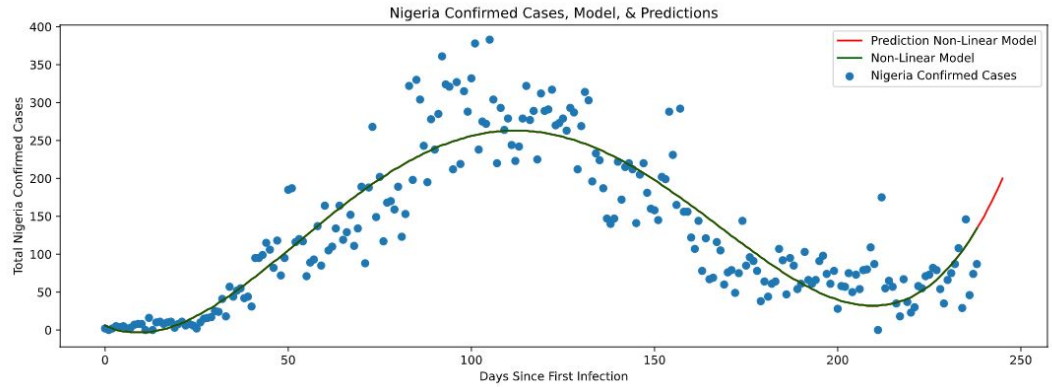


- Confirmed Cases: Mexico is predicted to see a slight increase of new Confirmed cases in the week of 11-13-2020 to 11-20-2020.

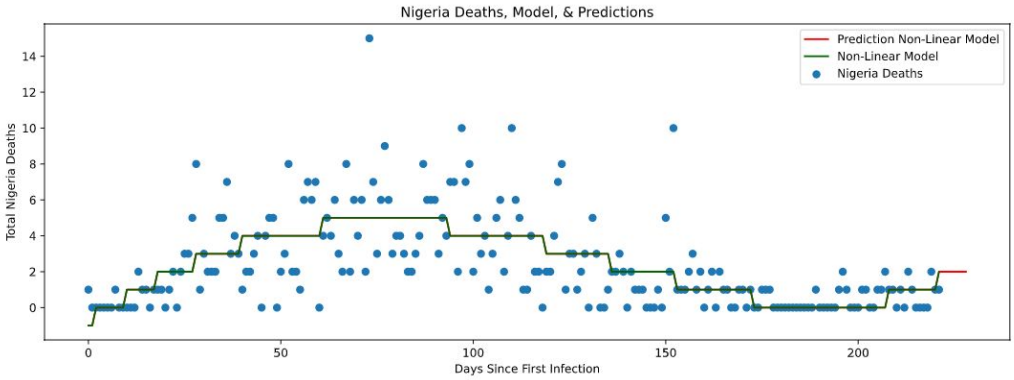
- Deaths: Mexico is predicted to see a slight increase of new deaths in the week of 11-13-2020 to 11-20-2020.

Nigeria

**** Normalize By Population Per 100,000,000 ****



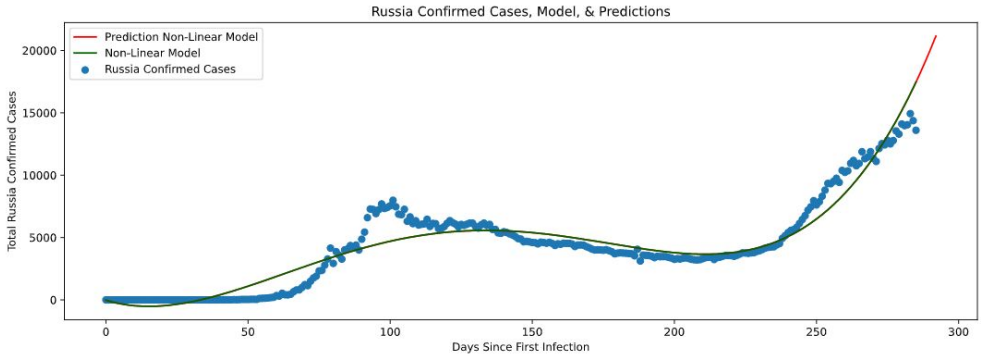
**** Normalize By Population Per 100,000,000 ****

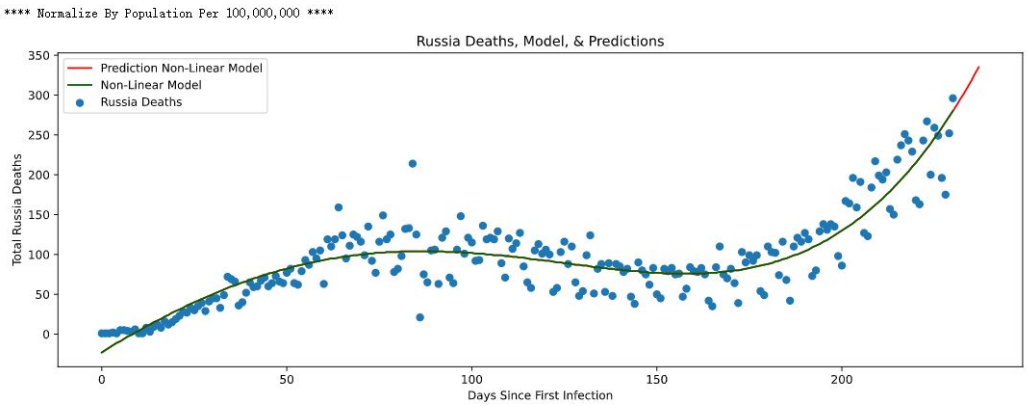


- Confirmed Cases: Nigeria is predicted to see an increase of new Confirmed cases in the week of 11-13-2020 to 11-20-2020.
- Deaths: Nigeria is predicted to see no change, hold at around 2 per 100,000,000 population, in new deaths in the week of 11-13-2020 to 11-20-2020.

Russia

**** Normalize By Population Per 100,000,000 ****





- Confirmed Cases: Russia is predicted to see a sharp increase of new Confirmed cases in the week of 11-13-2020 to 11-20-2020.
- Deaths: Russia is predicted to see an increase in new deaths in the week of 11-13-2020 to 11-20-2020.

V. DashBoard

Create a simple interactive dashboard to display the trends and predictions based on the analysis we have done so far.

References

USAFacts. *US Coronavirus Cases and Deaths*, https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/

United States Census Bureau. *ACS DEMOGRAPHIC AND HOUSING ESTIMATES*, https://data.census.gov/cedsci/table?q=dp&tid=ACSDP1Y2018.DP05

U.S. BUREAU OF LABOR STATISTICS. *Quarterly Census of Employment and Wages*, https://www.bls.gov/cew/downloadable-data-files.htm

Esri. *Definitive Healthcare: USA Hospital Beds*, https://coronavirus-resources.esri.com/datasets/1044bb19da8d4dbfb6a96eb1b4ebf629_0/data?geometry=-40.957%2C-16.820%2C15.996%2C72.123