SARS-CoV-2 Case Study

Advanced Data Science Capstone
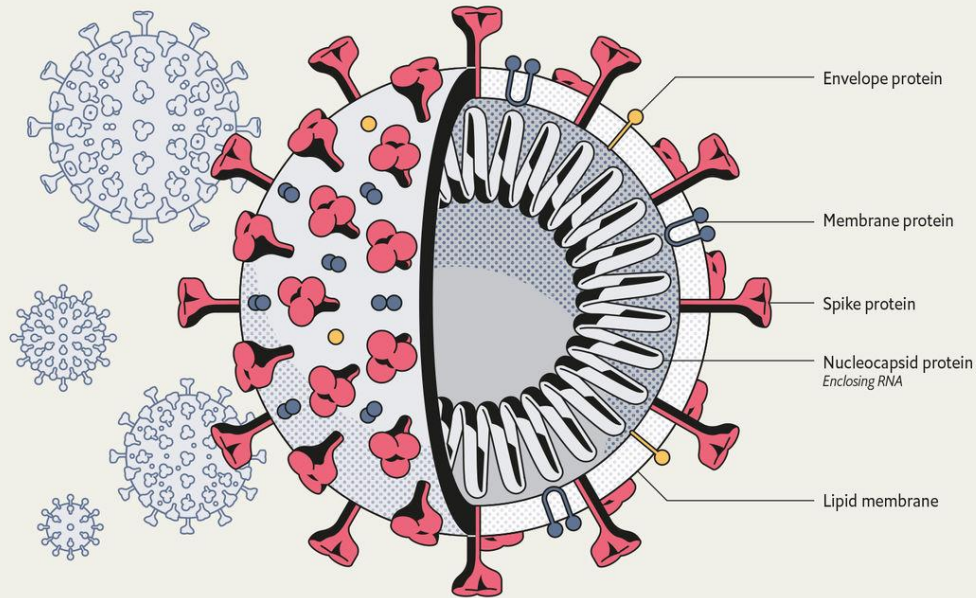
Submitted By: -

Sanuj Kumar

sanujsriv@gmail.com

Overview

- Background

- Motivation

- High level model

- Analysis
  - Spread Progression
  - Trend Projection
  - Daily Rise

- Models
  - Machine Learning
  - Deep Learning

- Visualizations

- Conclusion

Background

Labels on virus diagram:
- Envelope protein
- Membrane protein
- Spike protein
- Nucleocapsid protein *Enclosing RNA*
- Lipid membrane

Manuel Bortoletti

continent	Confirmed	Deaths	Recovered	Active	Incident_Rate	Mortality Rate (per 100)
Africa	721282	15169	380245	325868	3966.78	2.10
Asia	3348464	79140	2385810	883514	15840.23	2.36
Australia	13665	145	9976	3544	82.71	1.06
Europe	2660158	199052	1535157	854287	15503.29	7.48
North America	4439272	193803	1605861	2639608	4843.68	4.37
Others	15392	219	5860	9306	1971.80	1.42
South America	3251990	117634	2200818	933538	5718.37	3.62

o COVID-19 is the name of the "novel coronavirus" disease

o SARS-CoV-2 is the name of the virus that causes COVID-19

o COVID-19 is a new coronavirus disease

o Coronaviruses cause mild respiratory illnesses, such as the common cold
- o Severe Acute Respiratory Syndrome (SARS)
- o Middle East Respiratory Syndrome (MERS)
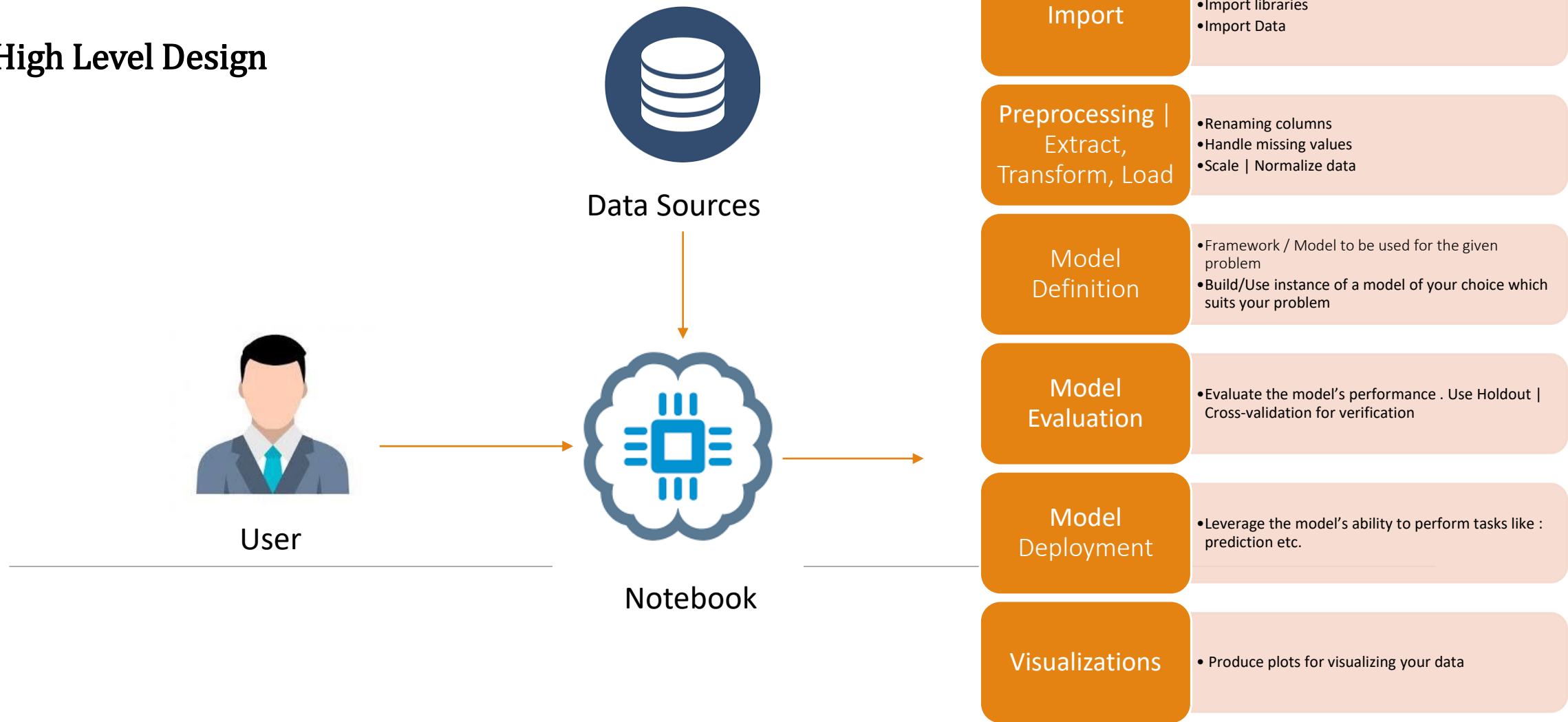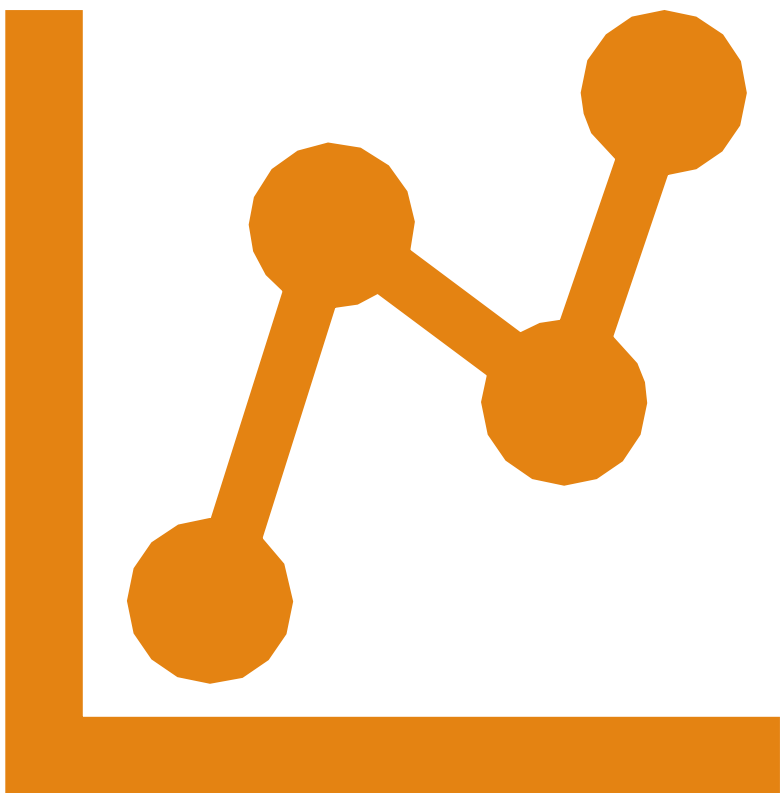
o Emerged from Hubei Province, China in December 2019

Image from : https://www.economist.com/img/b/1280/720/90/sites/default/files/20200314_FBD001.jpg

Motivation

" As the pandemic is spreading all over the world, it becomes more important to understand about this spread"

Confirmed	Deaths	Recovered	Active	Incident_Rate	Mortality Rate (per 100)
14538115	606922	8188292	5670485	48131.42	4.17

This project is an effort to analyze the spread progression, prediction of active cases, daily analysis of what is happening and visualization of cumulative data of confirmed, deaths, and recovered cases over time.

High Level Design

Data Sources

User

Notebook

Import
- Import libraries
- Import Data

Preprocessing | Extract, Transform, Load
- Renaming columns
- Handle missing values
- Scale | Normalize data

Model Definition
- Framework / Model to be used for the given problem
- Build/Use instance of a model of your choice which suits your problem

Model Evaluation
- Evaluate the model's performance . Use Holdout | Cross-validation for verification

Model Deployment
- Leverage the model's ability to perform tasks like : prediction etc.

Visualizations
- Produce plots for visualizing your data

Analysis

About Dataset

The data source is a repository from GitHub : https://github.com/CSSEGISandData/COVID-19.

This data repository is owned and operated by **Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE)** who are constantly working towards the cause i.e. analysis of COVID-19 and is deemed as a trustable source.

 It gets updated on a daily basis which makes it an ideal choice

Have a look at the dataset

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	1/28/20	1/29/20	1/30/20	1/31/20	2/1/20	2/2/20	2/3/20
0	NaN	Afghanistan	33.93911	67.709953	0	0	0	0	0	0	0	0	0	0	0	0	0
1	NaN	Albania	41.15330	20.168300	0	0	0	0	0	0	0	0	0	0	0	0	0

	Country_Region	Last_Update	Confirmed	Deaths	Recovered	Active	Delta_Confirmed	Delta_Recovered	Incident_Rate	People_Tested	People_Hospitalized	Province_State	FIPS	UID
0	Afghanistan	2020-01-22	0	0	NaN	NaN	0.0	NaN	0.0	NaN	NaN	NaN	NaN	4
1	Afghanistan	2020-01-23	0	0	NaN	NaN	0.0	NaN	0.0	NaN	NaN	NaN	NaN	4

	Country_Region	Last_Update	Lat	Long_	Confirmed	Deaths	Recovered	Active	Incident_Rate
0	Australia	2020-07-20 18:46:32	-25.0000	133.0000	12070.0	123.0	8395.0	3552.0	47.408257
1	Austria	2020-07-20 18:46:32	47.5162	14.5501	19743.0	711.0	17659.0	1373.0	219.210783
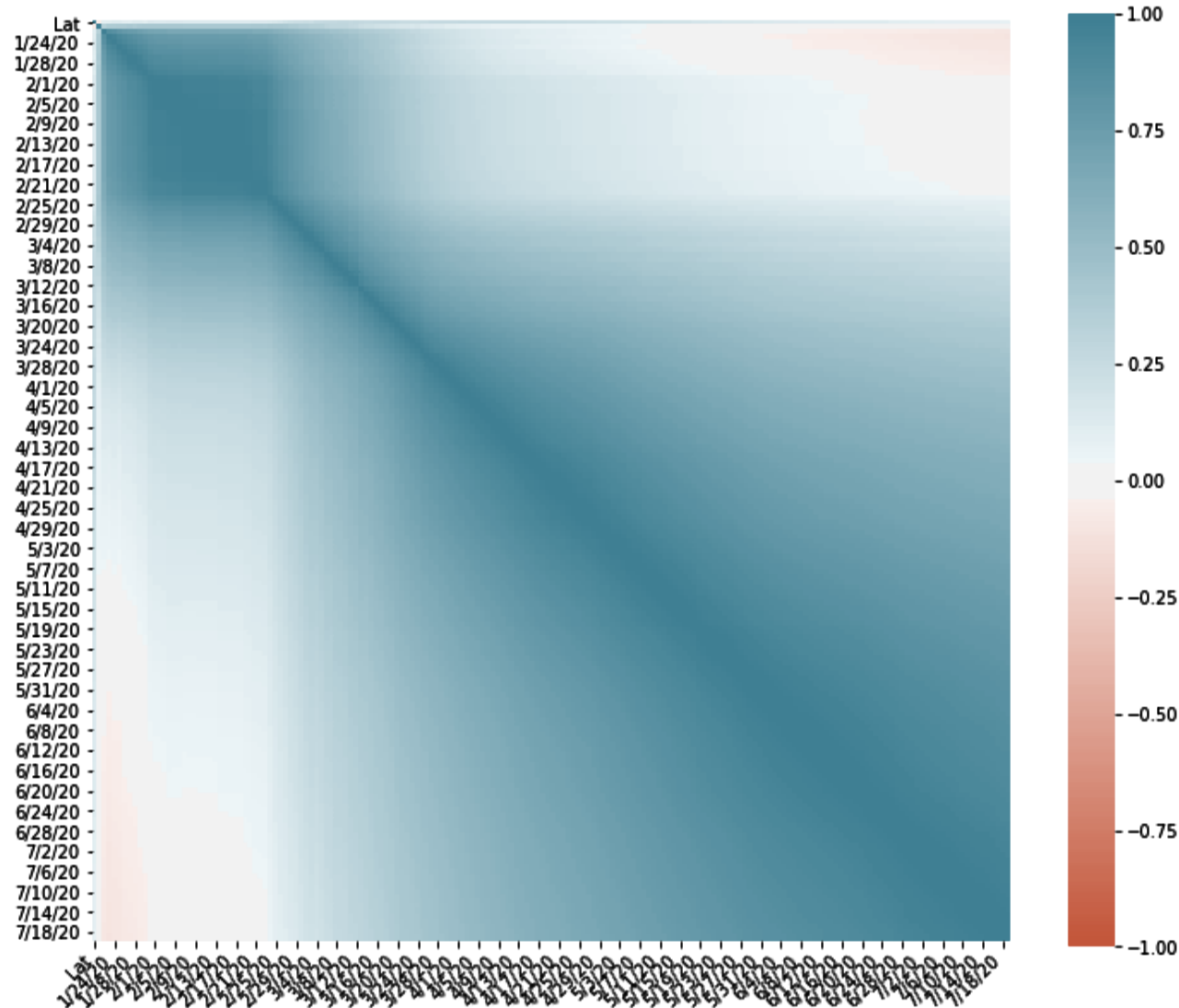
Preprocessing

 Data preprocessing – Scaling , Normalization , handling missing values, renaming , pruning, filling empty records

Correlation between the features was essential for data exploration.

After cleansing…
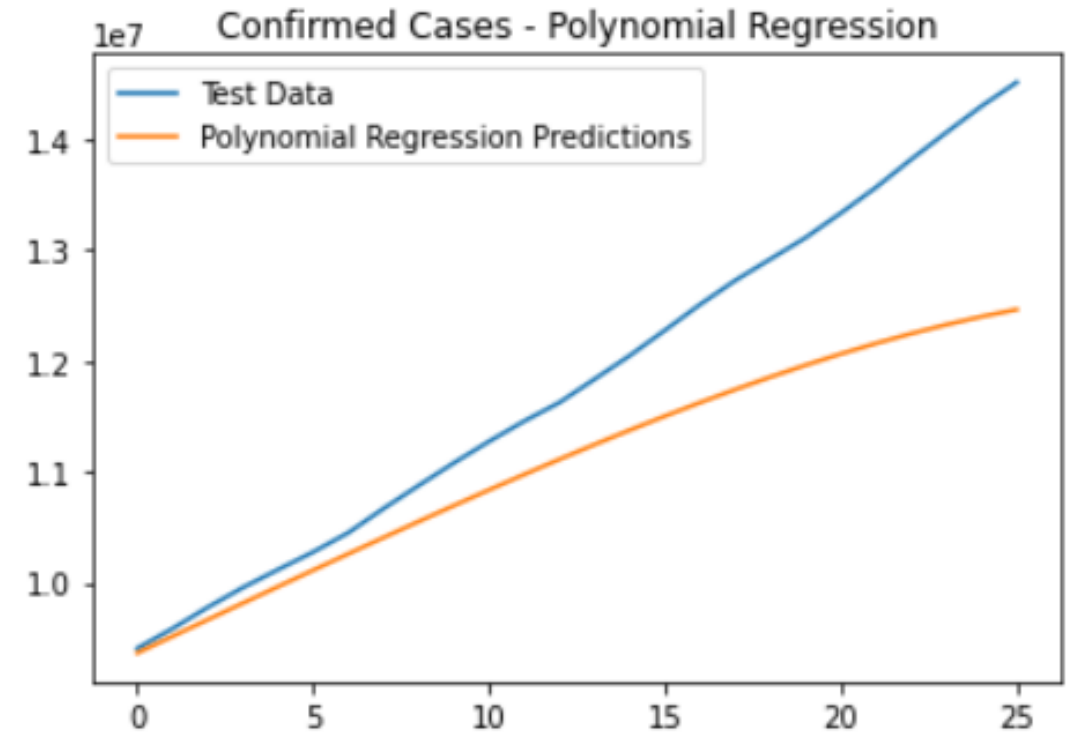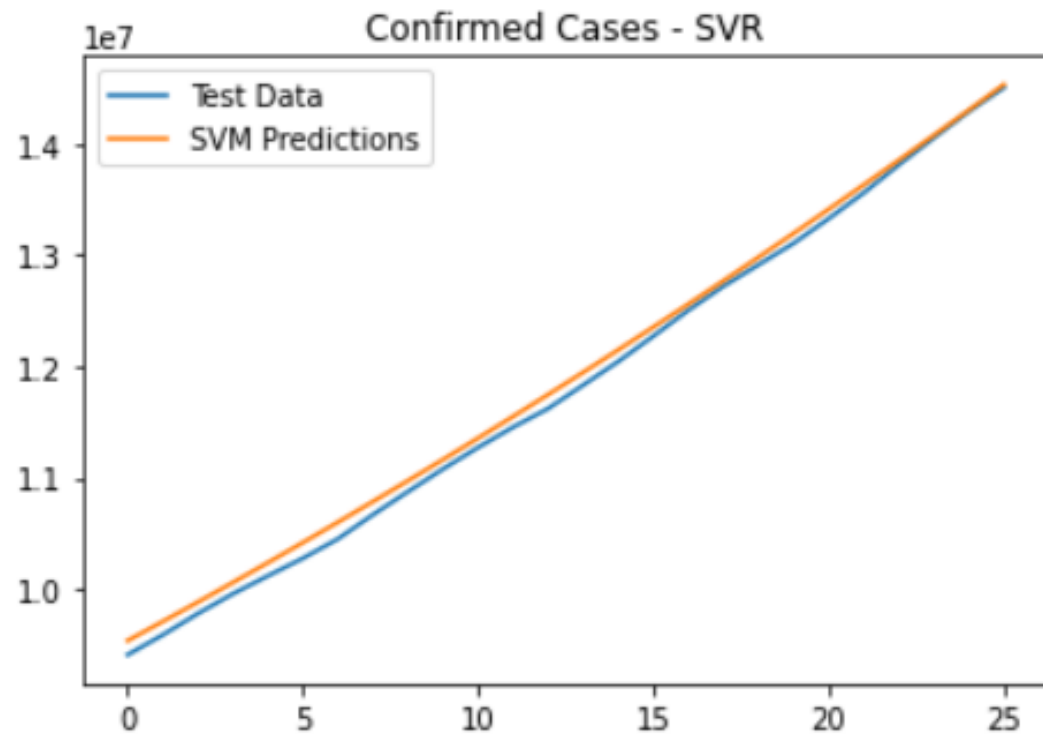
Models

Machine Learning Model

Python package Used : sklearn
Machine Learning Model used : Support Vector Regressor &
Polynomial Regression

The dataset I am using is highly dimensional which is suitable for models like Support Vector Machines & Polynomial Regression.

SVM is famous for efficiently performing a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

Also, the dataset which I am using grows daily making SVM and polynomial regression an ideal choice as they scale relatively well in these cases.

Confirmed Cases - SVR / Confirmed Cases - Polynomial Regression

Evaluation

Deep Learning Model

Framework Used: Keras

Keras a high-level deep learning framework that sits on top of backend frameworks like TensorFlow.

Keras is excellent because it allows you to experiment with different neural-nets with great speed! It sits atop other excellent frameworks like TensorFlow and lends well to the experienced as well as to novice data scientists! It doesn't require nearly as much code to get up and running!

Keras provides you with the flexibility to build all types of architectures; that could be recurrent neural networks, convolutional neural networks, simple neural networks, deep neural networks, etc.
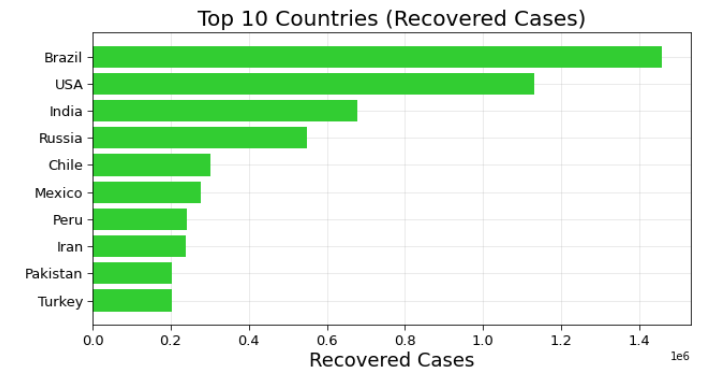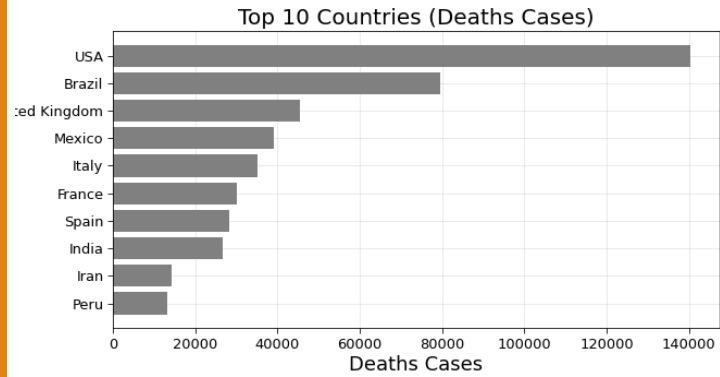
Evaluation
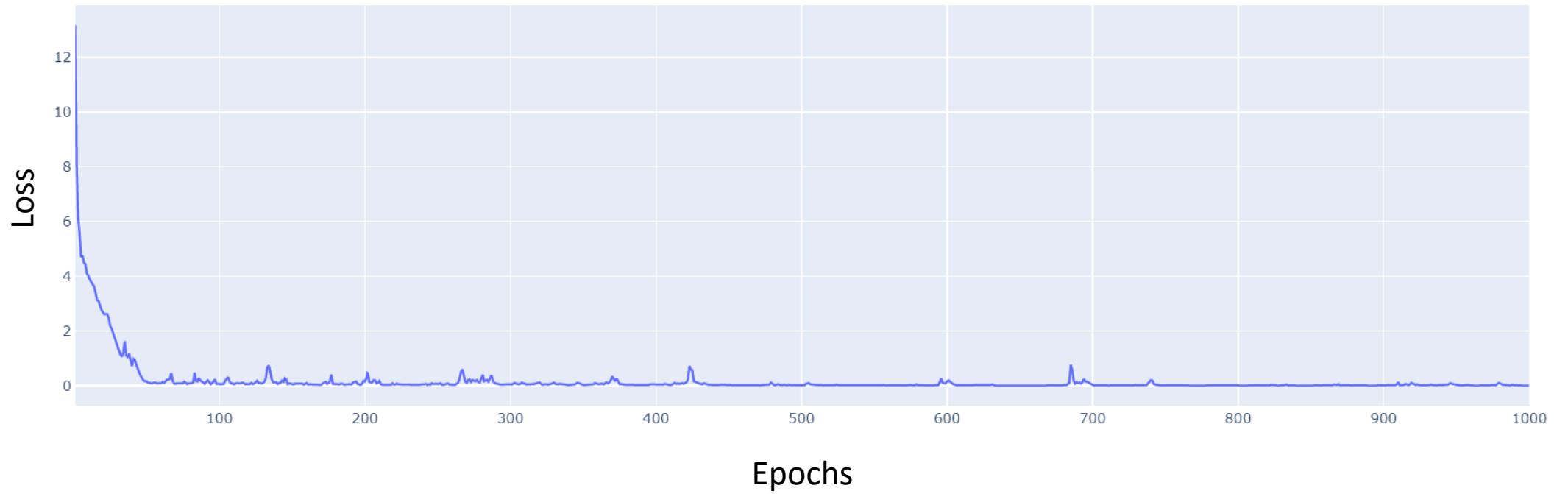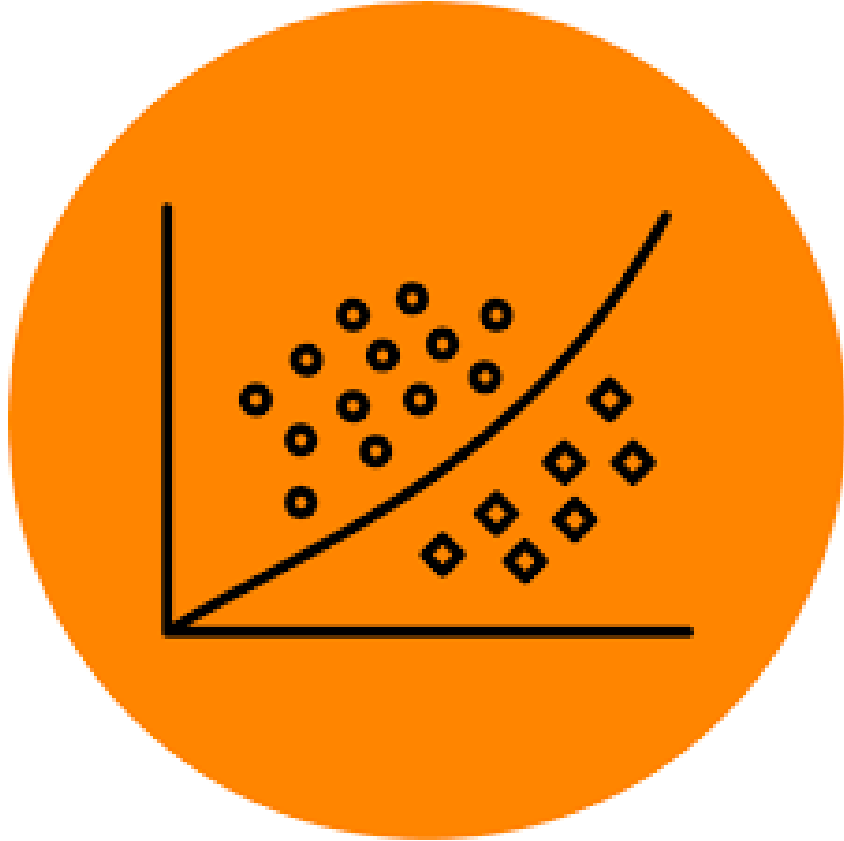
Visualization

World Trends
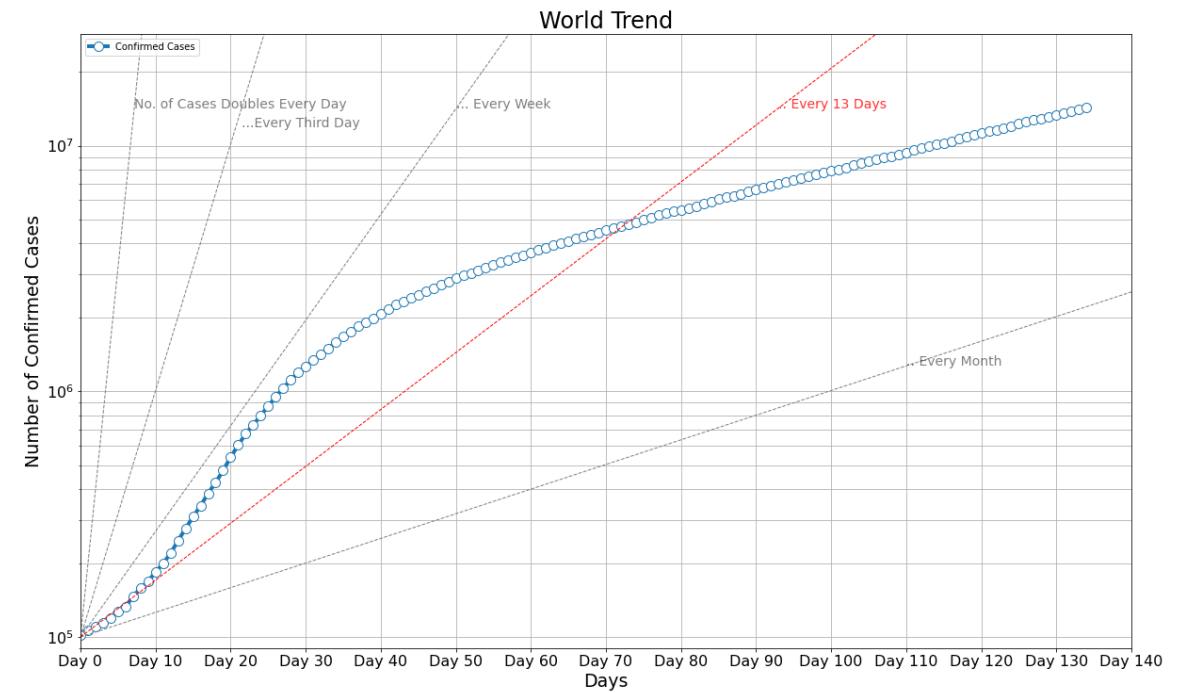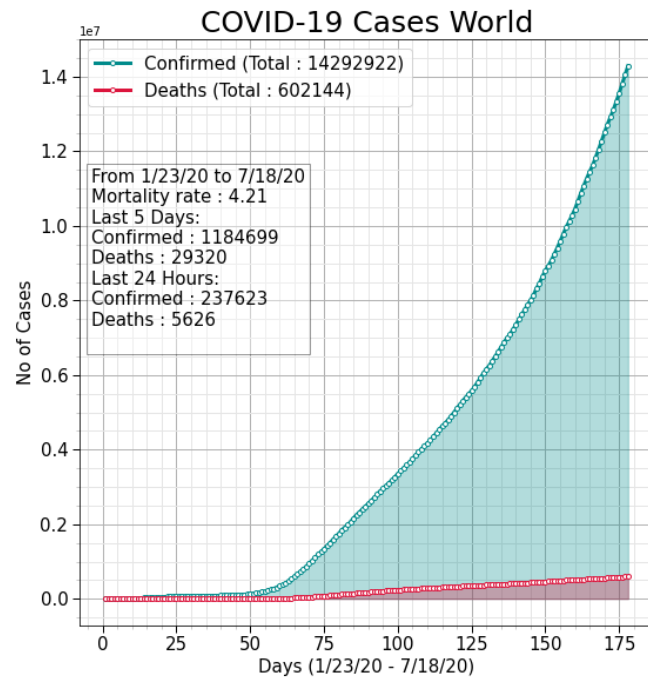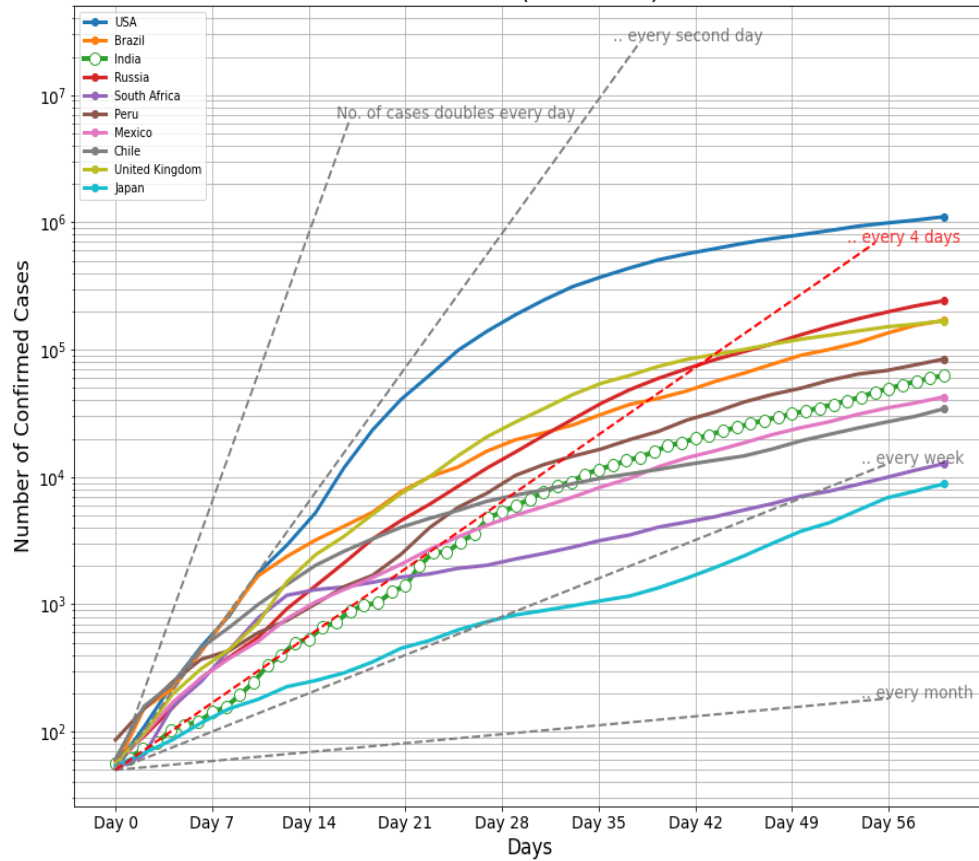
Trend
Comparison

Daily Cases

COVID-19 Global Daily New Confirmed Cases

COVID-19 Global Daily Deaths Reported
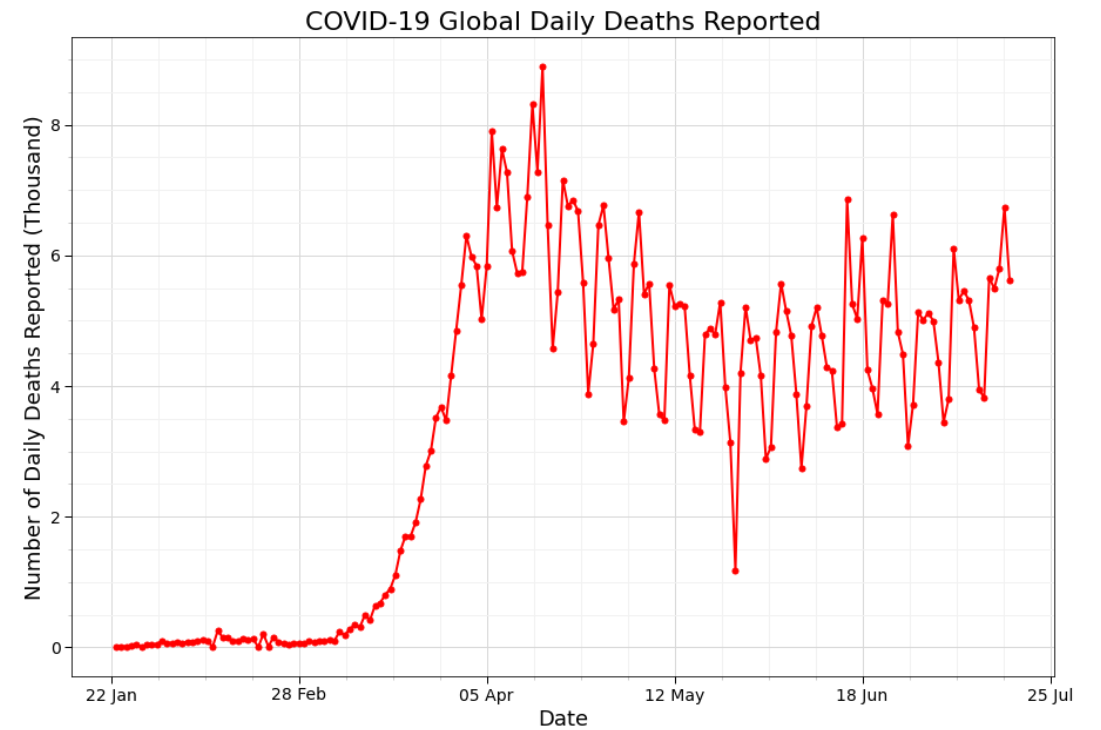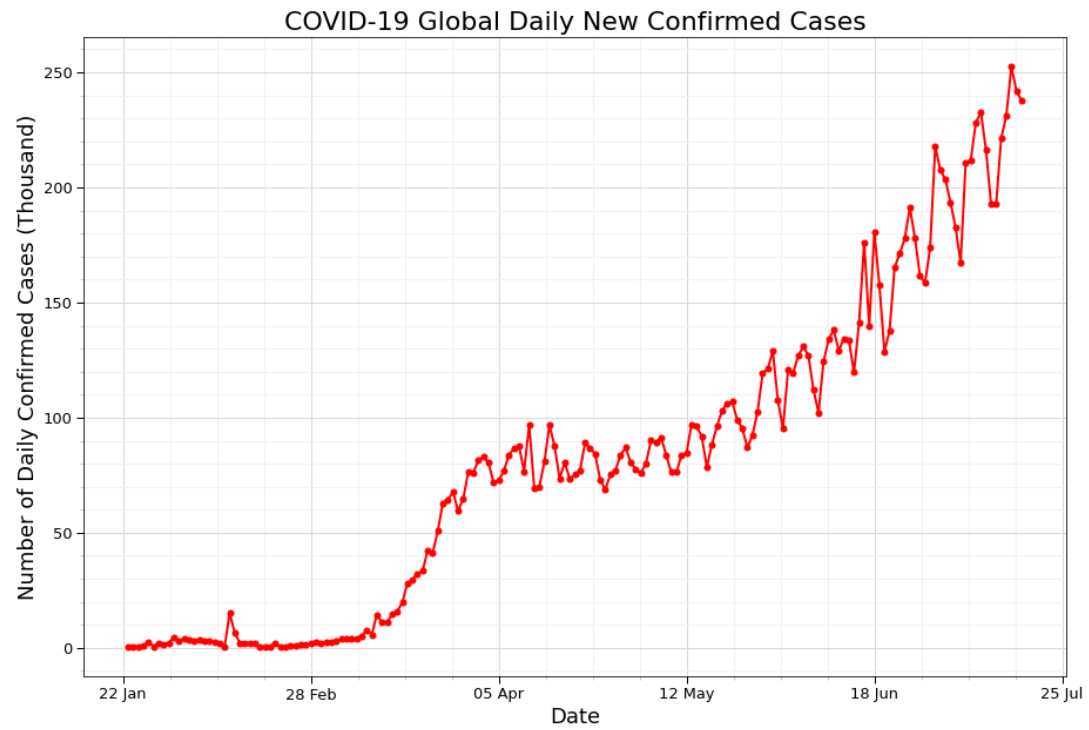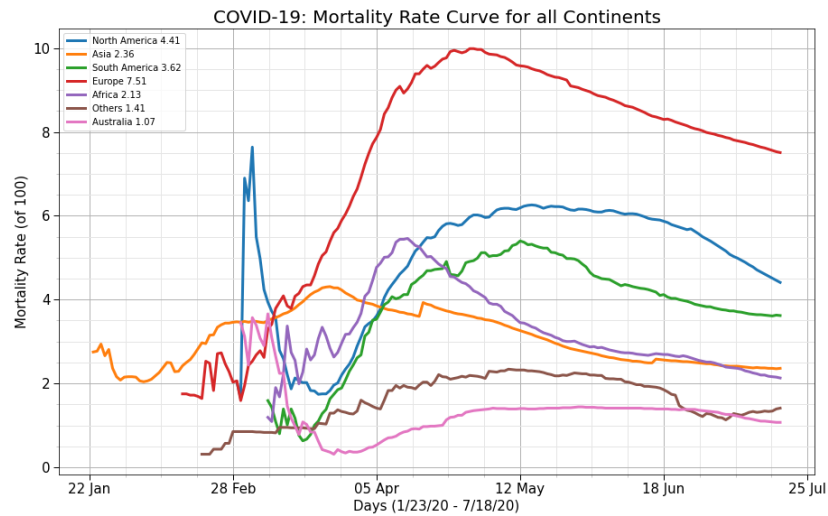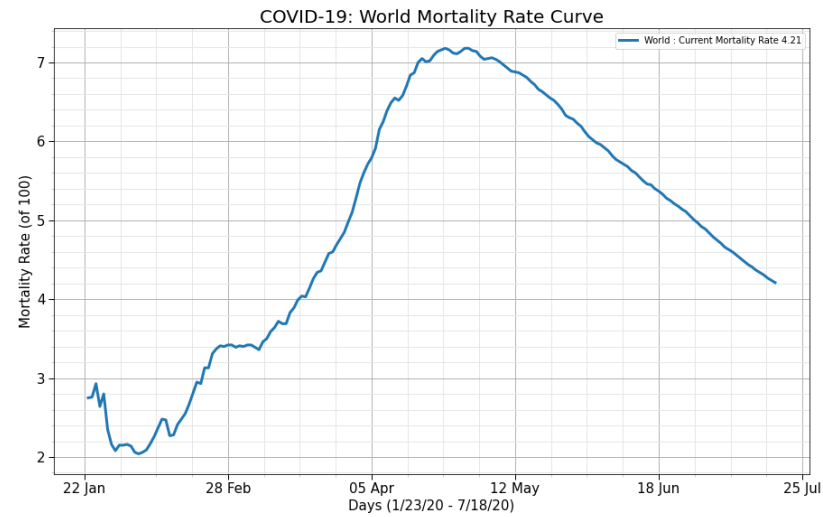
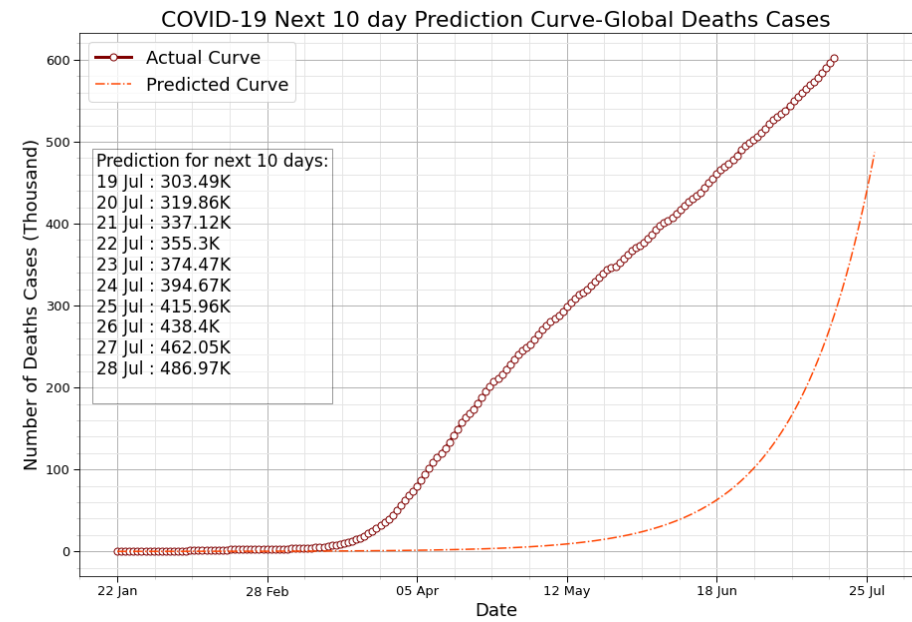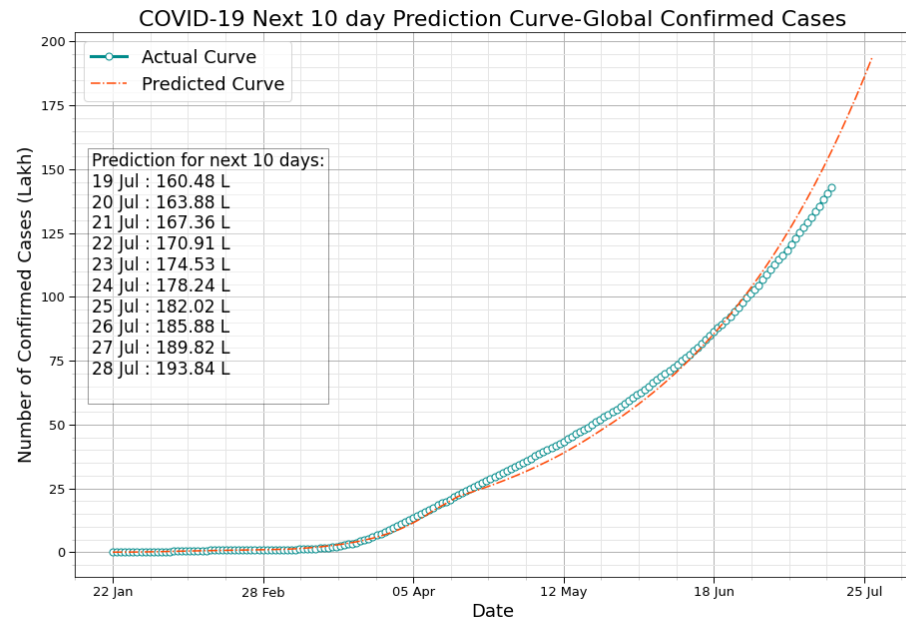# Mortality Rate
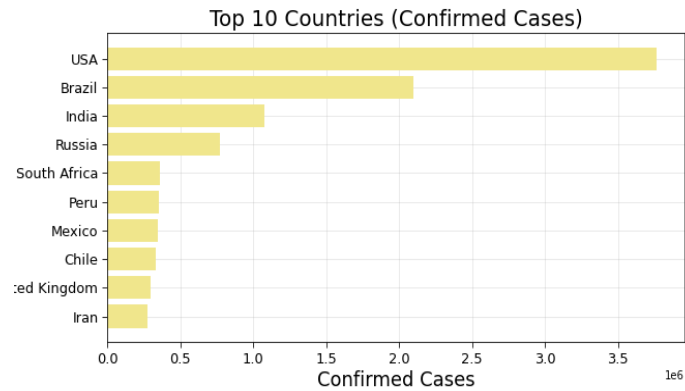
Top 10 Countries : Confirmed , Death and Recovered

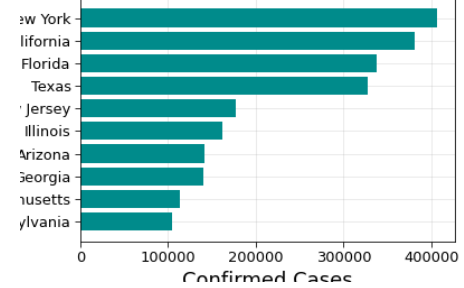COVID-19 Next 10 day Prediction Curve-Global Confirmed Cases

Prediction for next 10 days:
19 Jul : 160.48 L
20 Jul : 163.88 L
21 Jul : 167.36 L
22 Jul : 170.91 L
23 Jul : 174.53 L
24 Jul : 178.24 L
25 Jul : 182.02 L
26 Jul : 185.88 L
27 Jul : 189.82 L
28 Jul : 193.84 L

COVID-19 Next 10 day Prediction Curve-Global Deaths Cases

Prediction for next 10 days:
19 Jul : 303.49K
20 Jul : 319.86K
21 Jul : 337.12K
22 Jul : 355.3K
23 Jul : 374.47K
24 Jul : 394.67K
25 Jul : 415.96K
26 Jul : 438.4K
27 Jul : 462.05K
28 Jul : 486.97K

Prediction Curve

USA Covid-19 Analysis

Conclusion

Everyone wants to turn the page on COVID-19 as soon as possible. But lessons are inevitable. And it's up to each of us to decide if those lessons are right.

Throughout mankind's long history, there have been many such cases but no one expected something like this to happen in the 21st century.

We have to stay strong and stay indoors as much as possible for our own sake and continue to act from a moral perspective. After all, our best bet is a happy future for all who live on Earth, our common home.