Discrete Mathematics

SHORT REPORT

Covid-19 Analysis and Prediction

MEMBERS

No.	Name	StudentID
1	Nguyen Nam Kha (leader)	2052515
2	Nguyen Ngoc Hoa	2052485
3	Duong Ngoc Quang Huy	2052489
4	Nguyen Hoang Anh Thu	2053478
5	Nguyen Hong Quan	2052228

TABLE OF CONTENTS

01

METHOD

We will use Python to implement the task of collecting data as well as building models. Python's large standard library, commonly cited as one of its greatest strengths, provides tools suited to many tasks. We will use some Python's libraries that can help us implement such tasks like Data collecting, Data analytics and Machine learning:

- **pandas** library offers data structures and operations for manipulating numerical tables and time series.
- **numpy** library has a large collection of high-level mathematical functions to operate on arrays and matrices.
- **matplotlib** library is used for creating interactive visualizations in Python like plotting graphs.
- **sklearn** library supports building prediction models by applying many mathematical functions, specifically Linear Regression and Polynomial Regression.

02

DATA COLLECTING

2.1 Get raw data

Our data will be collected from an organization called “Our World in Data (OWID)” and stored in a .csv file, we will get and read data from the link below:
https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/owid-covid-data.csv

2.2 Take useful features for analysing data

The features that we use for analyze:

- Location (to get the country name)
- Date
- New_cases

2.3 Fix NaN cells

NaN cells are cells that have missing data.
We fix them by replacing them with 0.

2.4 Get data of the specific object

To have an overview of the pandemic, we analyze the general data of the entire World. Therefore, we filter World in feature Location

2.5 Sort by month

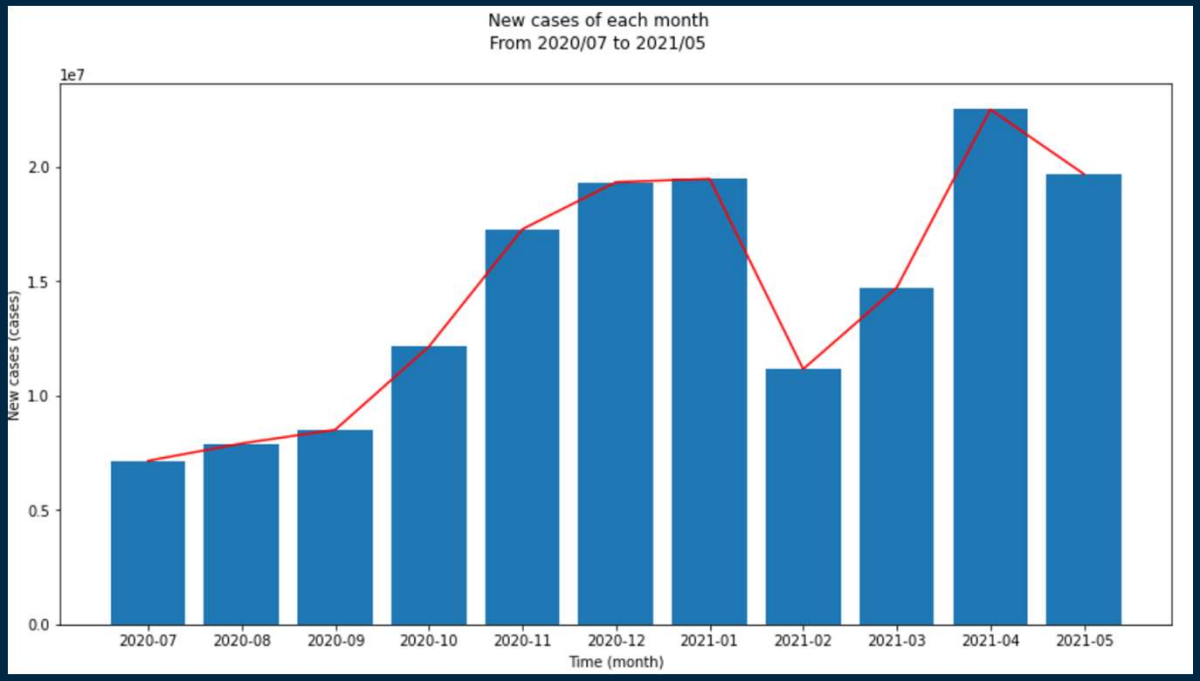
We take data from 2020/07 to 2020/12 (before we used the vaccine) and from 2021/01 to 2021/05 (started using vaccine)

03

DATA ANALYSIS

3.1 Plot the data

First we have to plot the data by drawing a mixed chart – a combination of a bar chart and a line graph.



3.2 Analyze the data

- From July 2020 to December 2020, the tendency seemingly resembles a normal distribution in which two conditions remained unchanged – the spread of the virus and the quick calls for lockdowns worldwide.
- In the first two months of 2021, January and February, the tendency follows an exponential distribution. An abrupt fall in figure is a clear indicator of the effects of vaccination and lockdowns at that period.
- In the next two months, the trajectory of a gamma distribution with K= 7,5 and theta=1.0 was found due to the drastic growth of number. This sudden surge may result from lagging vaccination campaigns worldwide and too-soon reopens in multiple countries.
- May 2021 presented a declining trend, indicating that more lockdowns and successful vaccinations had occurred.

04

MODEL PREDICTION

Linear Regression and Polynomial Regression

4.1.1 Theory

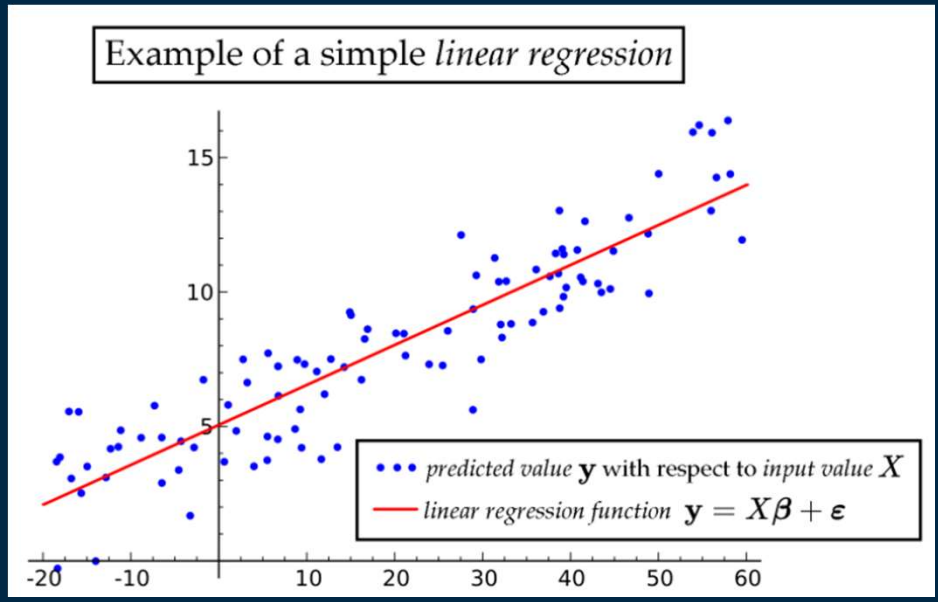
Linear Regression model :

- In statistics, linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).
- Given a data set of n statistical units, a linear regression model assumes that the relationship between the dependent variable y and the p-vector of regressors x is linear with the apperance of error variable ε — an unobserved random variable that adds "noise" to the linear relationship between the dependent variable and regressors.
- The model takes the form :

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

4.1.1 Theory (cont.)

So the theory of linear regression is very simple:
We have to find the regression coefficient so that the red-line function best fits the blue-dot value, which means the error term is minimized. After that, we will use this red-line function to predict the desired value.



4.1.1 Theory (cont.)

Polynomial Regession model :

- Polynomial regression is quite similar to linear regression. The difference between them is the relationship between the independent variable **X** and the dependent variable y is modelled as an nth degree polynomial in **X**. For example:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon.$$

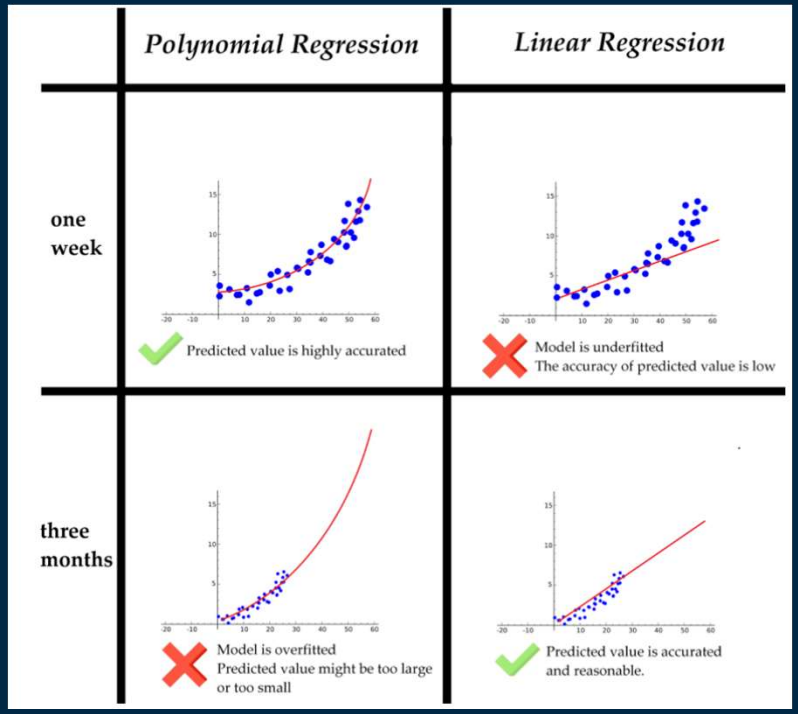
- The function above is the quadratic model of the form. Polynomial regression fits a nonlinear relationship between the value of X and the corresponding conditional mean of y. Therefore, we can yield a general function of polynomial regression model:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon.$$

4.1.1 Theory (cont.)

We will use *polynomial regression* for the one-week period. However for the case of three-month period, we have to use *linear regression*, here's the reason why:

- Using *linear regression* to predict such a short period like one-week, the model might be underfitted and our predicted value might be inaccurate.
- Applying *polynomial regression* on a three-month prediction might cause the model to be overfitted, which means our predicted value will be impractical.



4.1.2 Create DoMath function

This function receives 3 parameters:

- The training sets (x,y)
- Number of days after 1 week and 3 months (ow_norm, tm_norm)
- Identified parameter (check)

And it will return 4 parameters:

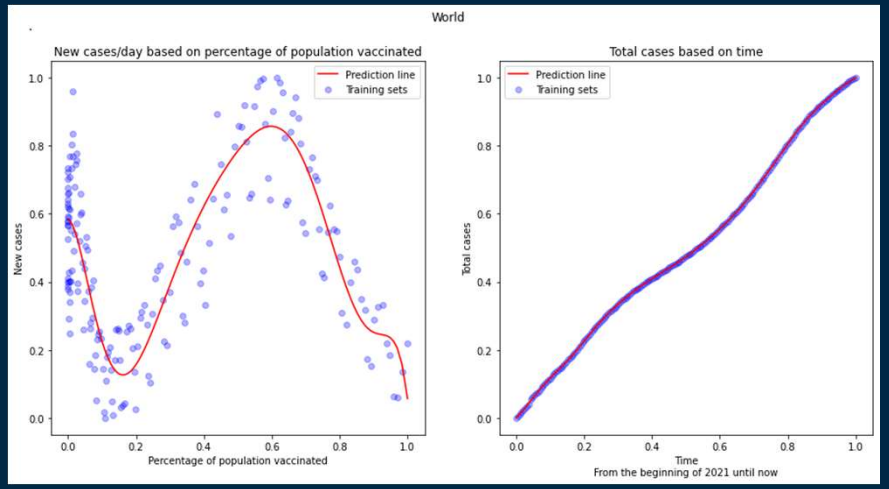
- Polynomial Regression function (poly_re)
- Linear Regression function (lin_re)
- Total cases predicted after 1 week and 3 months (ow_re, tm_re)
- Accuracy of the model (Best Accuracy)

Global

4.2.1 Building model

Now we will build a model to predict COVID-19 situation of the World.

Step 1: Get data from open source, clean data.
Step 2: Normalize the data.
Step 3: Create training sets.
Step 4: Create input.
Step 5: Perform math.
Step 6: Plot the graph.
Step 7: Prediction and calculate Accuracy.



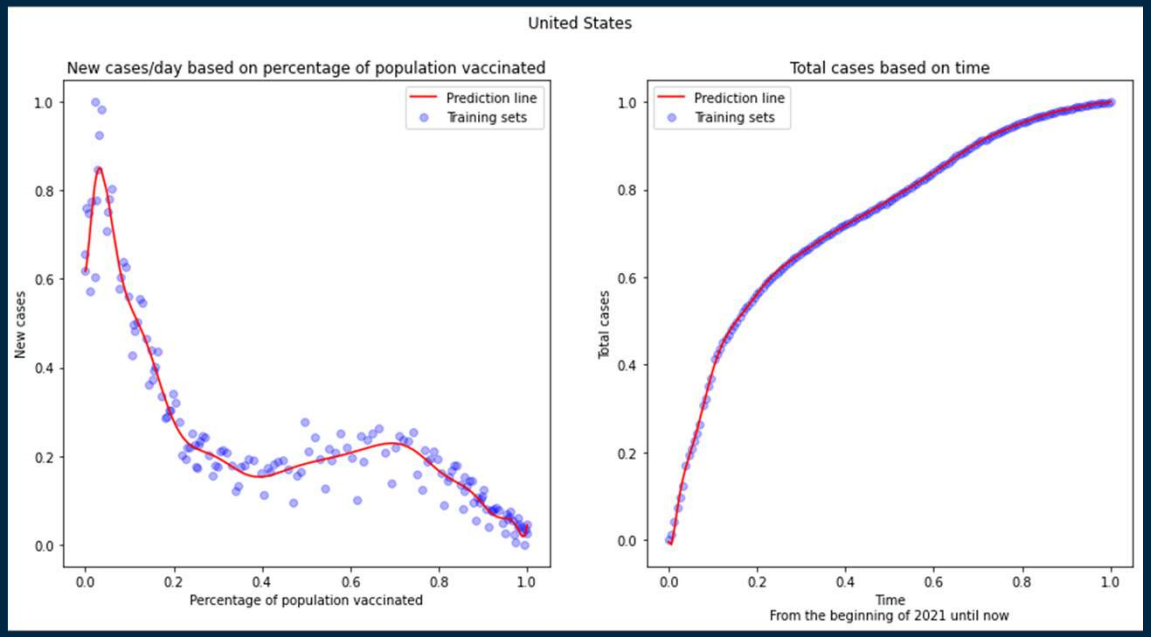
3.2.2 Create Prediction function

We will combine all the steps above into one function that can be reused several times to predict the pandemic situation around the world. This function is called Prediction.

Other countries

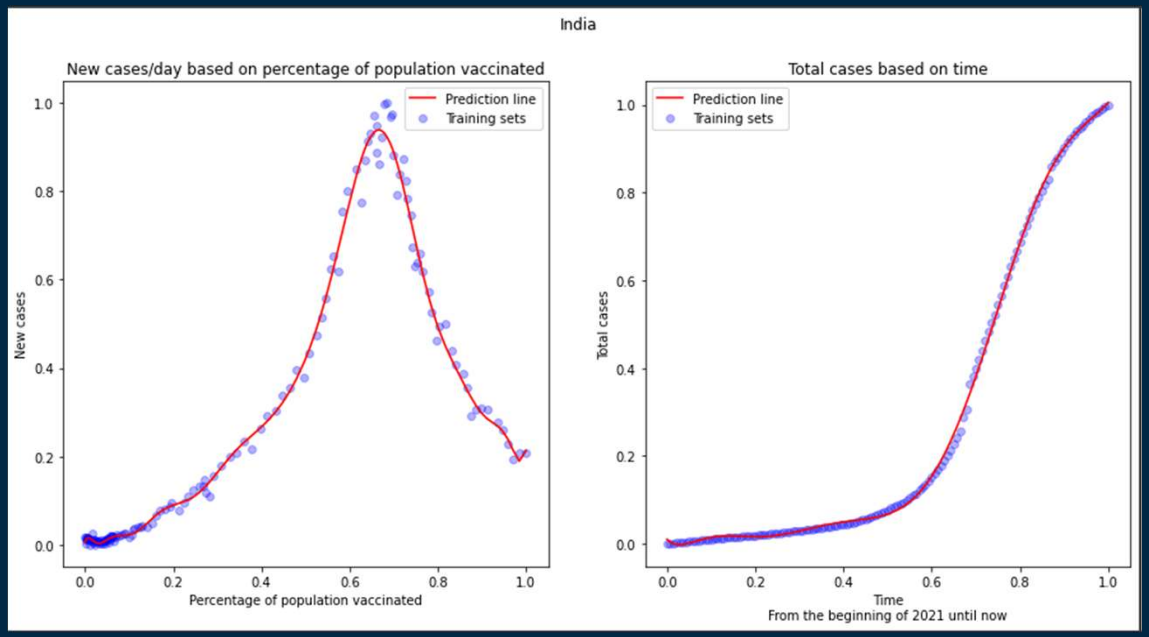
4.3.1 USA, India, Brazil

After one week, total cases of United States will reach: 34812918 cases
After three months, total cases of United States will reach: 42090080 cases
Accuracy of this prediction model: 99.9474%



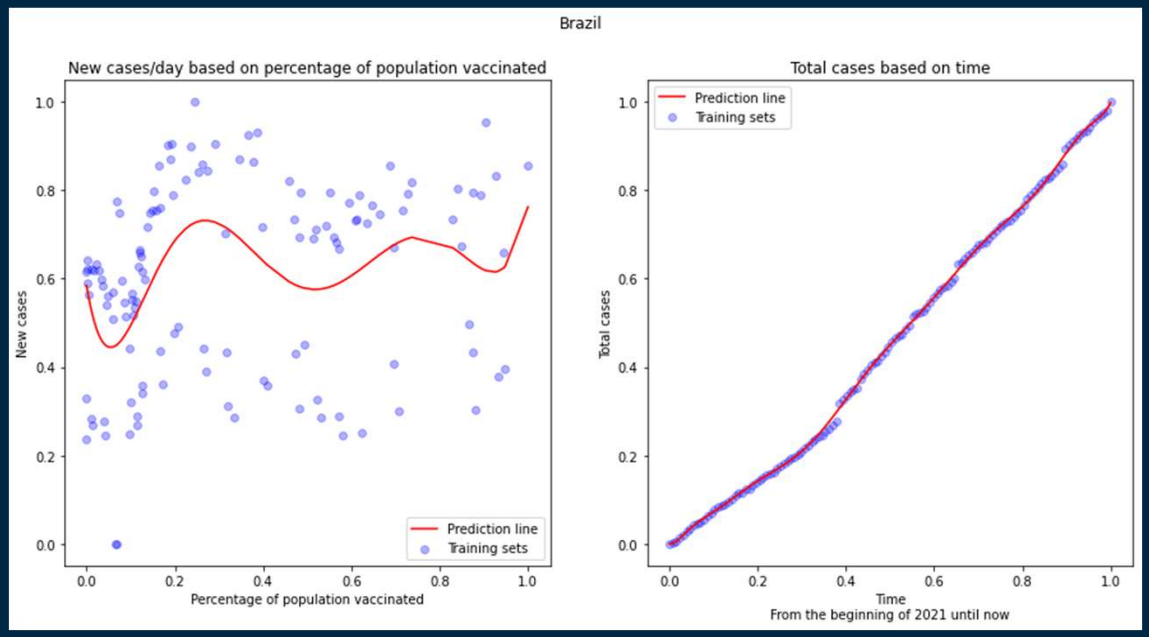
4.3.1 USA, India, Brazil (cont.)

After one week, total cases of India will reach: 30775092 cases
After three months, total cases of India will reach: 38903814cases
Accuracy of this prediction model: 99.9152%



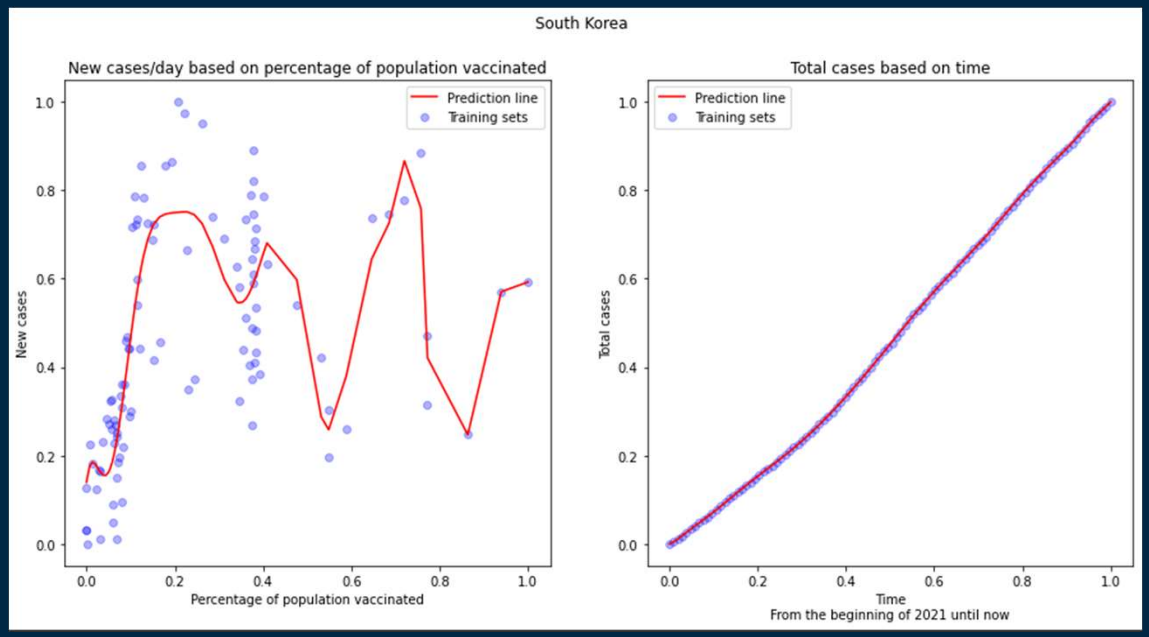
4.3.1 USA, India, Brazil (cont.)

After one week, total cases of Brazil will reach: 22656820 cases
After three months, total cases of Brazil will reach: 23275536 cases
Accuracy of this prediction model: 99.9241%



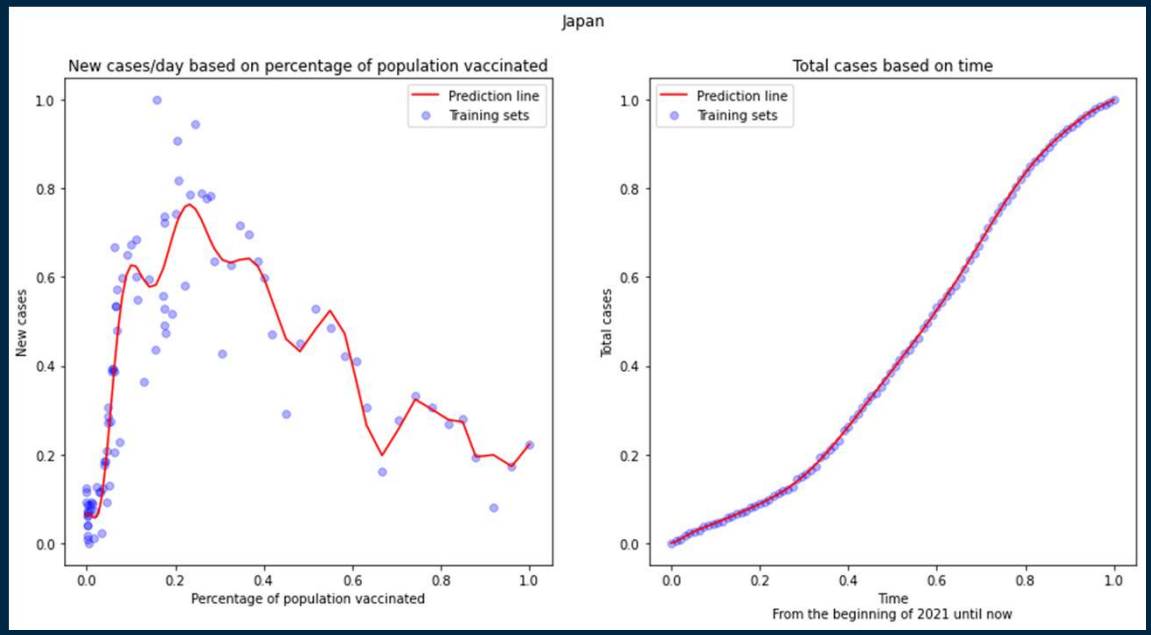
4.3.2 Korea, Japan, Vietnam

After one week, total cases of Korea each: 190391 cases
After three months, total cases of Korea will reach: 198341 cases
Accuracy of this prediction model: 99.9951%



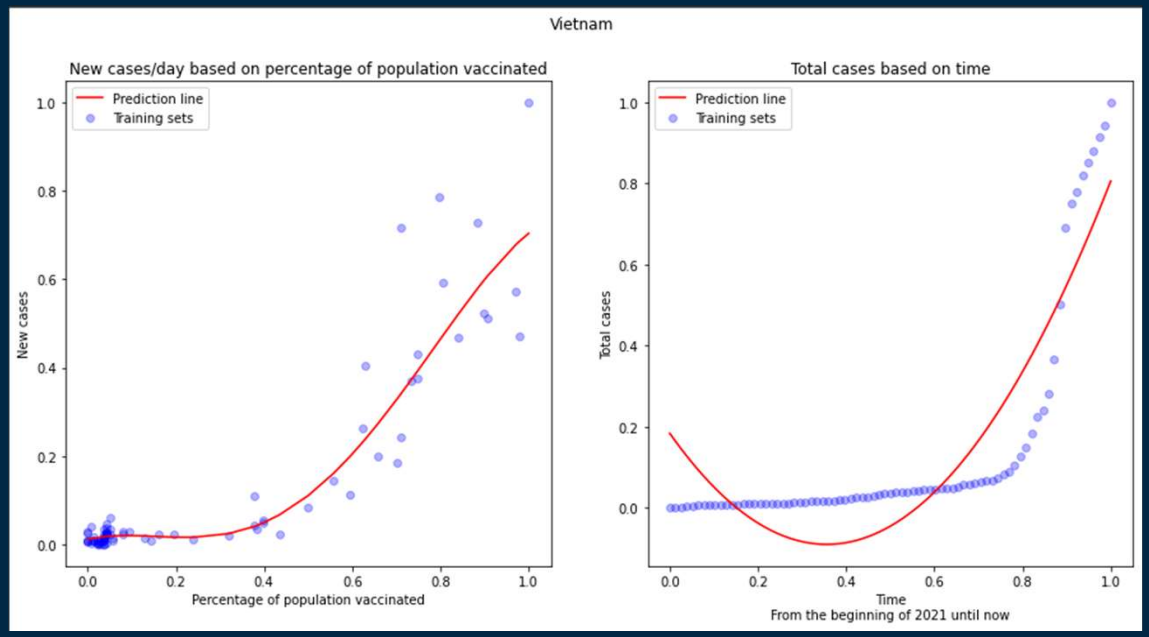
4.3.2 Korea, Japan, Vietnam (cont.)

After one week, total cases of Japan will reach: 1106450 cases
After three months, total cases of Japan will reach: 1153635 cases
Accuracy of this prediction model: 99.9895%



4.3.2 Korea, Japan, Vietnam (cont.)

After one week, total cases of Vietnam will reach: 10450 cases
After three months, total cases of Vietnam will reach: 11335 cases
Accuracy of this prediction model: 70.9482%



05

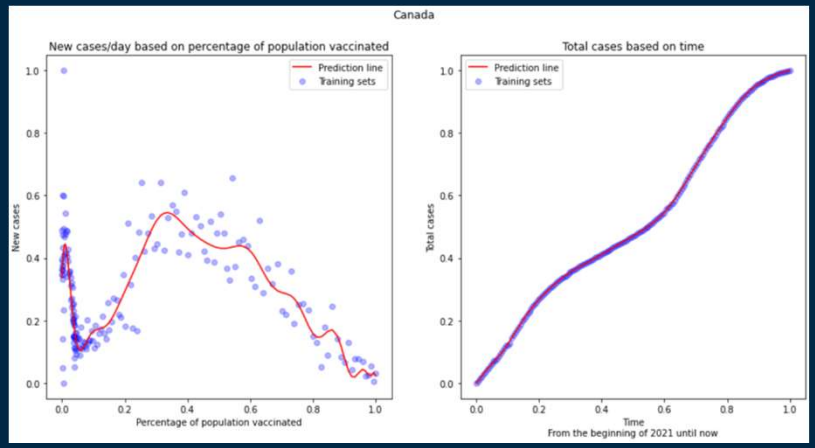
MODEL ANALYSIS

5.1 Ability

- Our model is built based on Linear Regression and Poly Regression.
- Through the graph that illustrates number of new cases per day based of percentage of people vaccinated, we can know the pandemic situation before and after vaccination, the trend (increasing or decreasing) in the spread of the pandemic at the current
- Through the graph that illustrates the number of total cases based on time, we can predict the pandemic situation a week and three months later, thereby knowing the growth rate and the peak of the pandemic.

5.2 Flexibility

- Our model operates by collecting data of 4 features: total cases, new cases, percentage of population vaccinated and time so we can give predictions to almost every country in the world.
- However, this model cannot be used for the prediction of some countries that don't publish their data or lack one of the above features.
- For example, our model can predict the pandemic situation in Canada :



Advantages and Disadvantages

Advantages

Linear Regression:

- Simple to implement.
- Gives information about the relevance of features.
- Performs well when the dataset is linearly separable.
- Works well irrespective of the dataset size (thanks to the normalization process)
- Overfitting can be reduced by using some dimensionality reduction techniques, such as regularization.

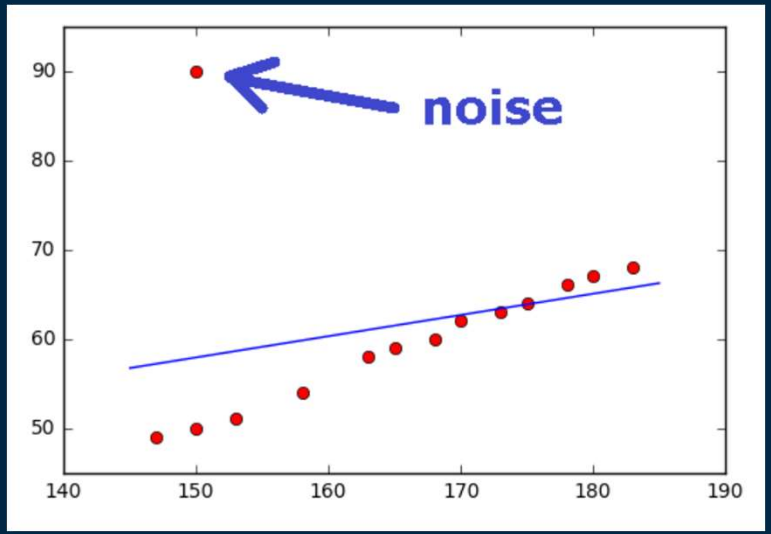
Polynomial Regression:

- Provides a wide range of functions that best fit the relationship between the dependent and independent variable.

Disadvantages

Linear Regression:

- Oversimplifies the problems since in the real world, data rarely has a linear relationship between dependent and independent variables.
- Assumes there is a linear relationship (a straight line) between dependent and independent variables
- Really prone to underfitting, specifically it is very sensitive to noise.



Polynomial Regression:

- Very sensitive to noise and easy to get overfitted.

Choosing the right model

- Linear regression model is a good choice to predict the total cases of one specific country, because the number of total cases always increases.
- Polynomial regression is a better choice to model the relationship between new cases per day and percentage of population vaccinated, since the dependent value, which is new cases per day, does not always increase or decrease.
- Polynomial regression is suitable to make predictions for a short period ahead of us such as one week later.
- Linear regression is a better choice to predict the pandemic situation three months later since it will reduce the possibility of overfitting.

THANKS