An: màu cam

Vinh: màu xanh

Nghiệp: màu lục

Đại: màu đỏ

COVID-19 DATA ANALYTICS AND FORECASTING (temporary name)

**First part: Beginning sections that every report has**

1. ***Introduction (done)***
2. ***Tasks***

Step 0: Collect the data

Step 1: Exploratory data analysis (EDA) a.k.a Data profiling - to get a deeper understanding of our data. The subtasks are listed below:

*a) Data cleaning*: check for duplicate rows, missing data, N/A cells, and (if possible) wrongly recorded data. (cont.)

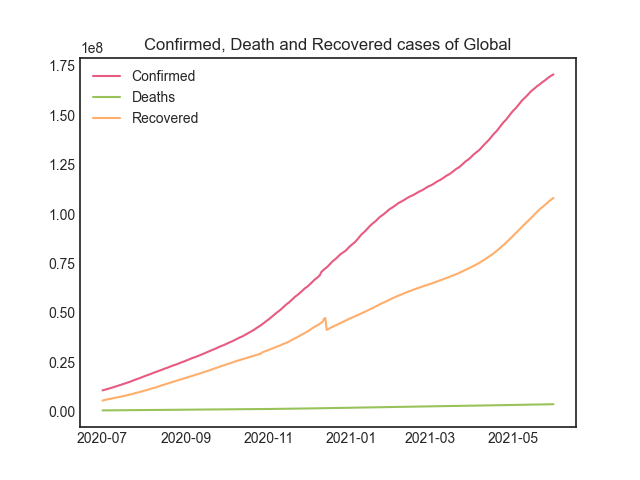
**Second part: COVID-19 terminology 101**

1. ***Introduction (done)***
2. ***Overview of epidemic***
3. ***Stages of Disease (done)***
4. **History of Pandemics** (done)
5. **The novel coronavirus (done)**

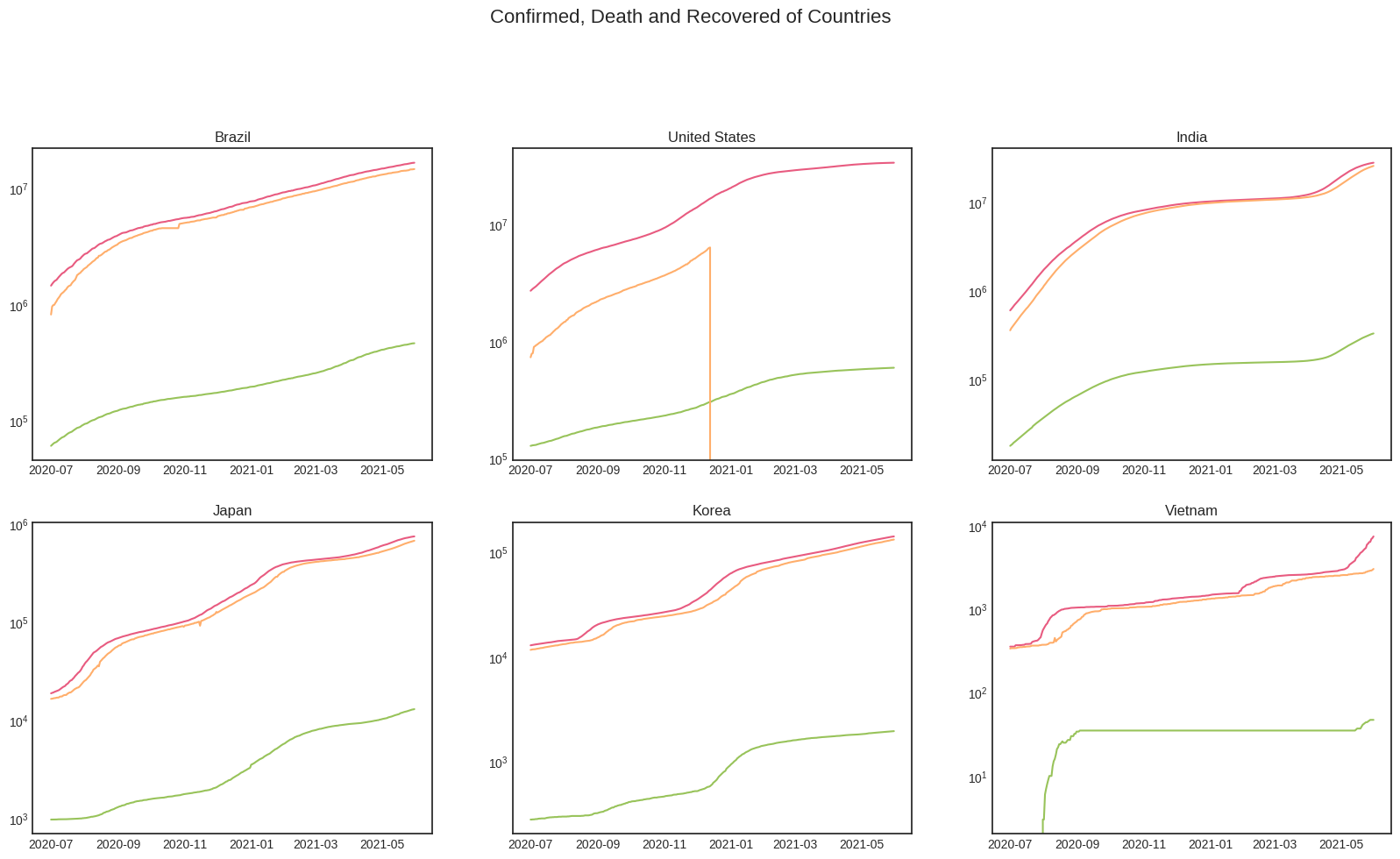
**Third part: EDA**

*EDA insights:*

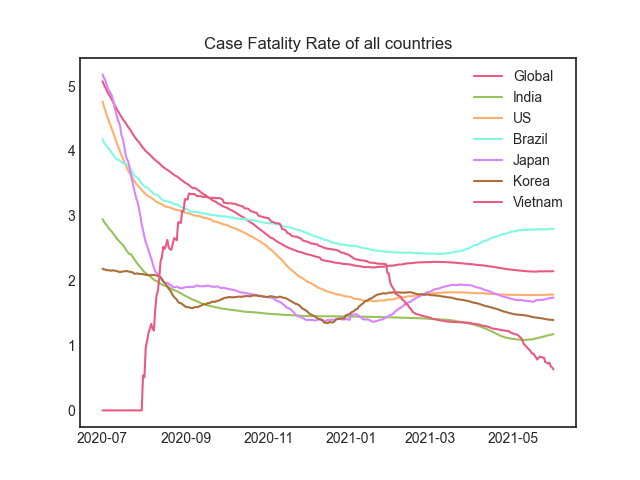
1. *Confirmed (C), Death (D) and Recovered* ® *of Global*

**

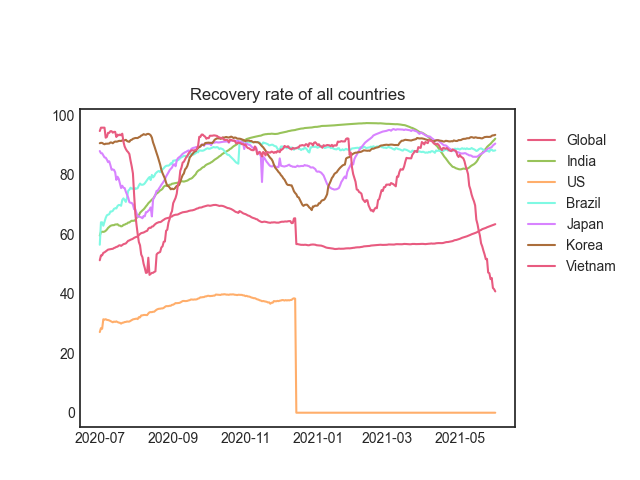
1. *C, D, R of each country (done)*

**

1. *Case-fatality rate (Death as a proportion of cases) of 6 countries with respect to Global (CFR = (Number of Death Cases / Number of Confirmed Cases) x 100)*

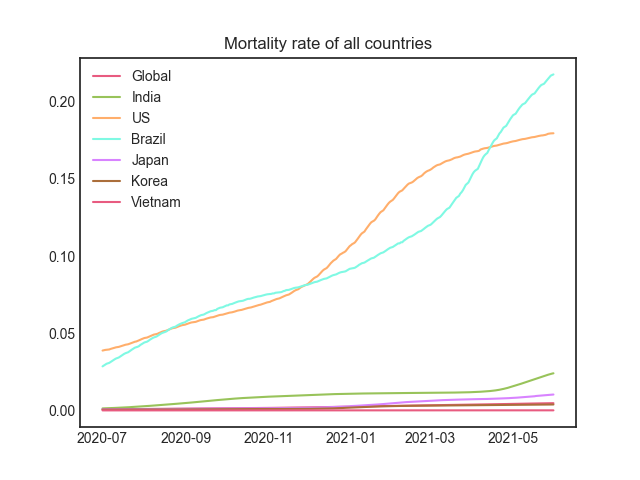
**

1. *Recovery rate (Recovered as a proportion of cases) of 6 countries with respect to Global (Recovery Rate= (Number of Recovered Cases / Number of Confirmed Cases) x 100)*

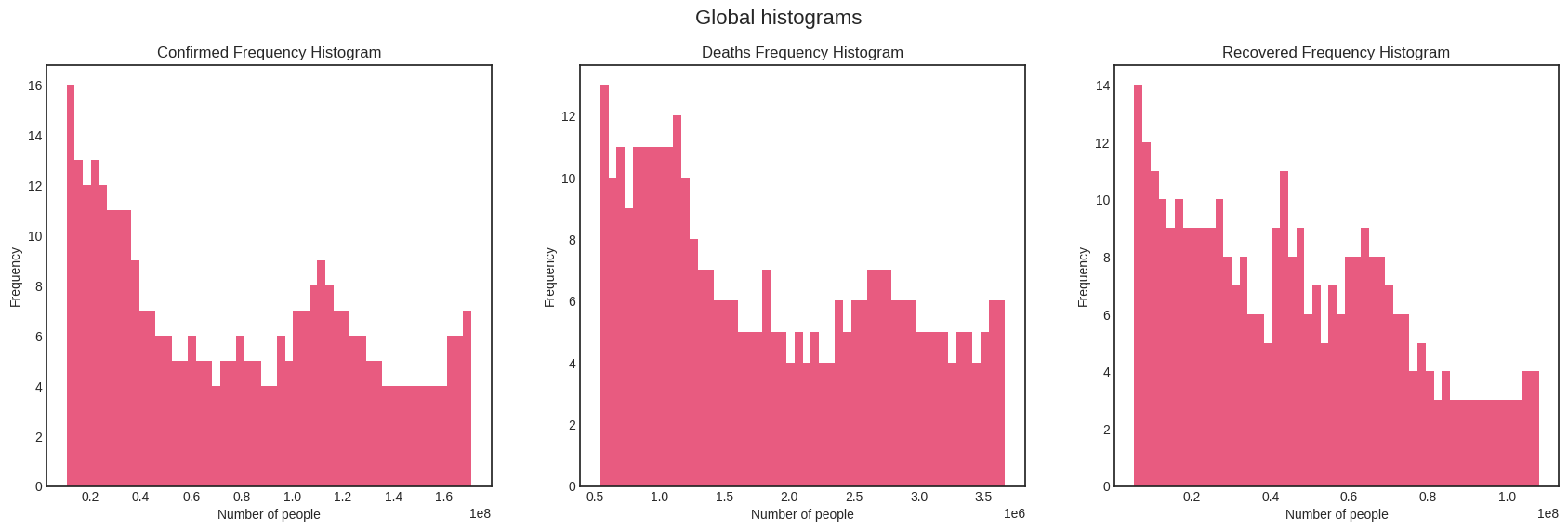
**

1. *Mortality rate (Death as a proportion of population) of 6 countries with respect to Global*

*(MR = Number of Death Cases / Population) x 100)*

**

1. *Distribution histogram for Global (done)*

**

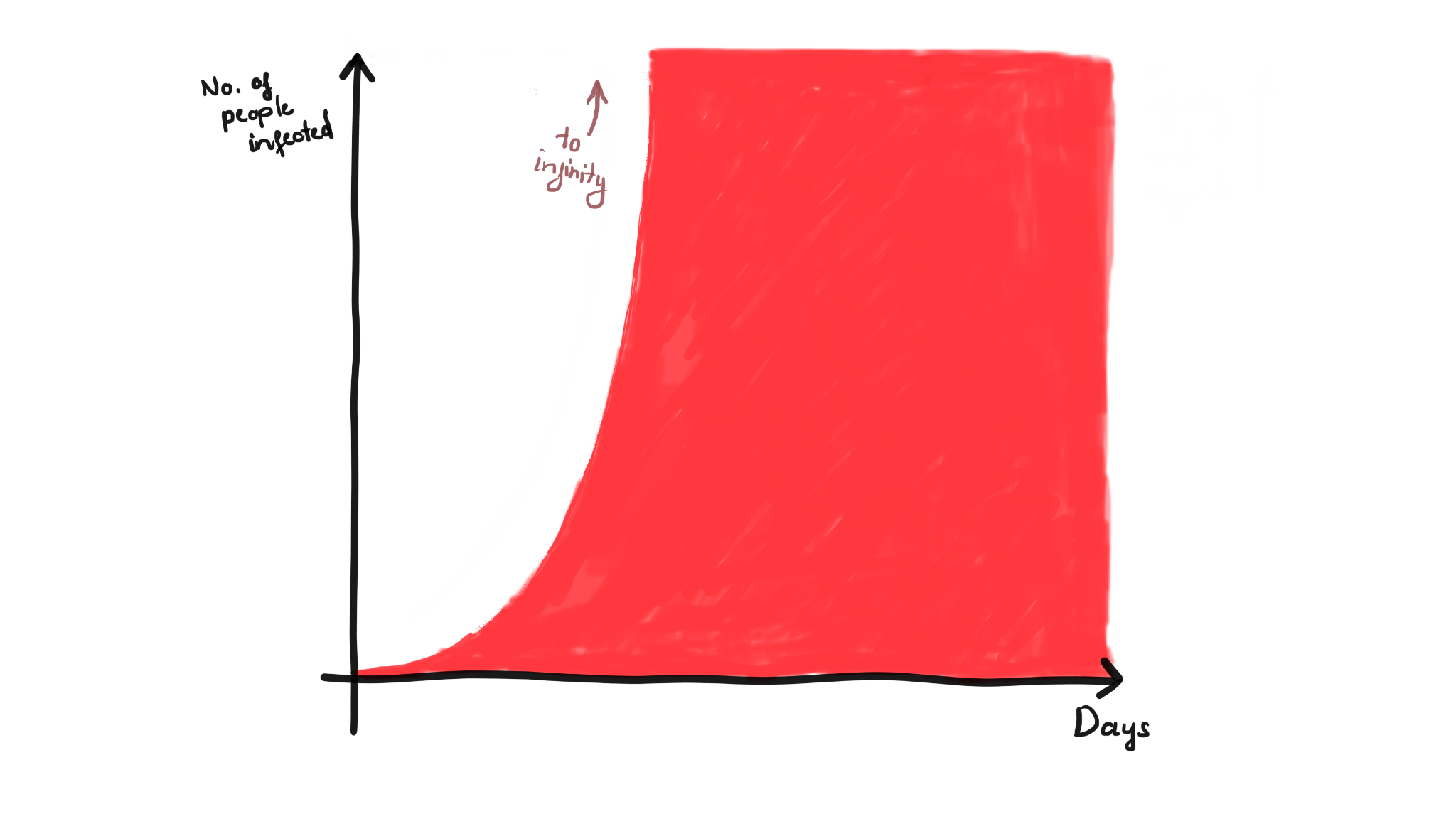
1. *Distribution histogram for Brazil*
2. *Distribution histogram for US*
3. *Distribution histogram for India*
4. *Distribution histogram for Japan*
5. *Distribution histogram for Korea*
6. *Distribution histogram for VN*
7. *Age distribution histogram (7 subplots inside one)*
8. *Stringency Index (6 subplot inside one)*

**Fourth part: SIR terminology 101**

1. ***From basic exponential growth to the famous curves set***

A person who contains the virus is called **Infected**, and others who are not infected but still have a percentage of chance to catch the virus spread by the Infected is called **Susceptible**. **Infectious** people can turn **Susceptible** people into more **Infectious** people.

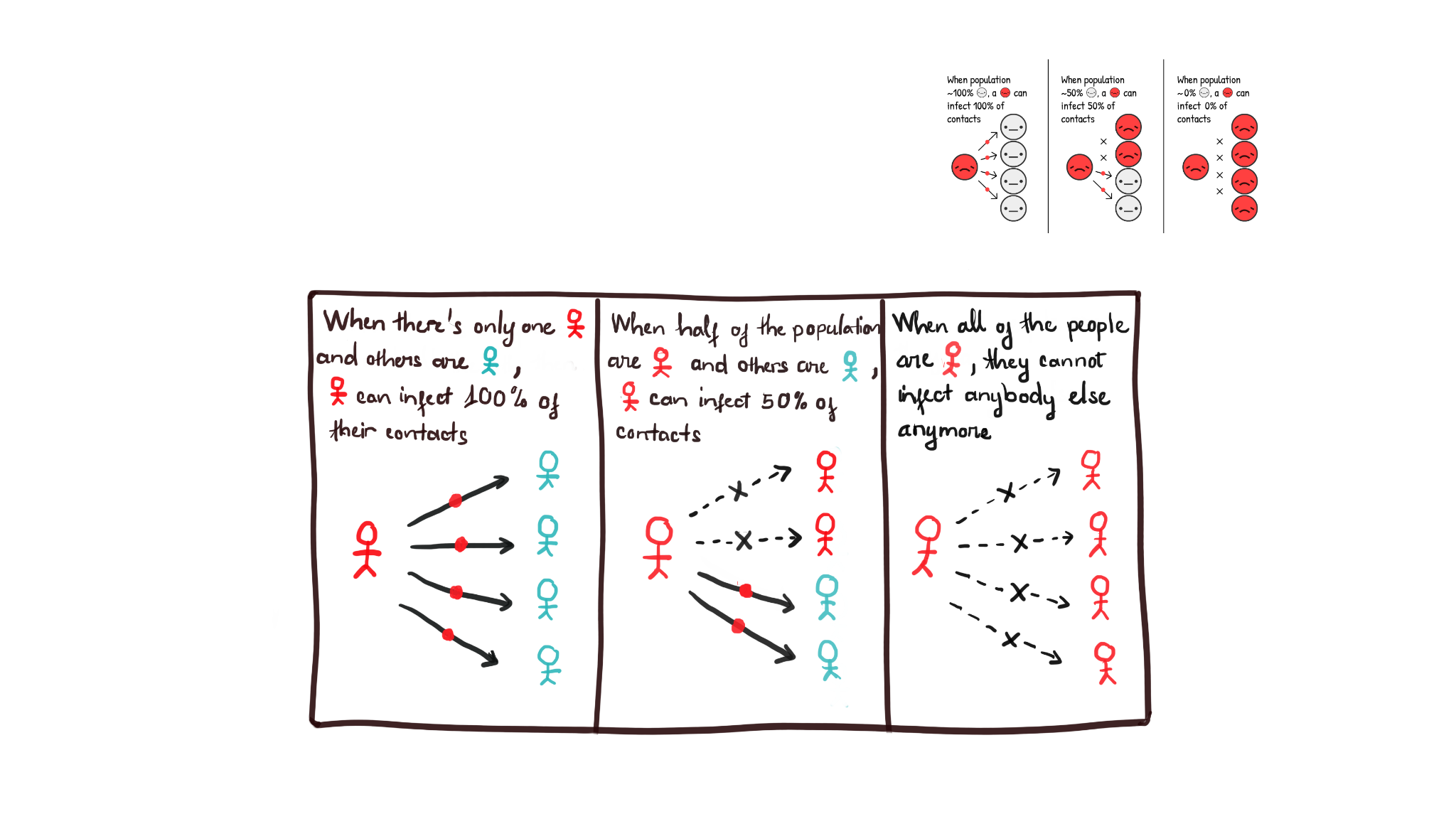
In a very simple scenario, for example, let’s say each **Infected** can infect two **Susceptible**, means that after some fixed interval of time, 1 **Infected** will become 2, 2 will become 4, 4 will become 8...and so on; starting with only 1 of the population being **Infected** and without any extra [yếu tố tác động], here is what the infection graph would look like:



[SIR Terminology-1]

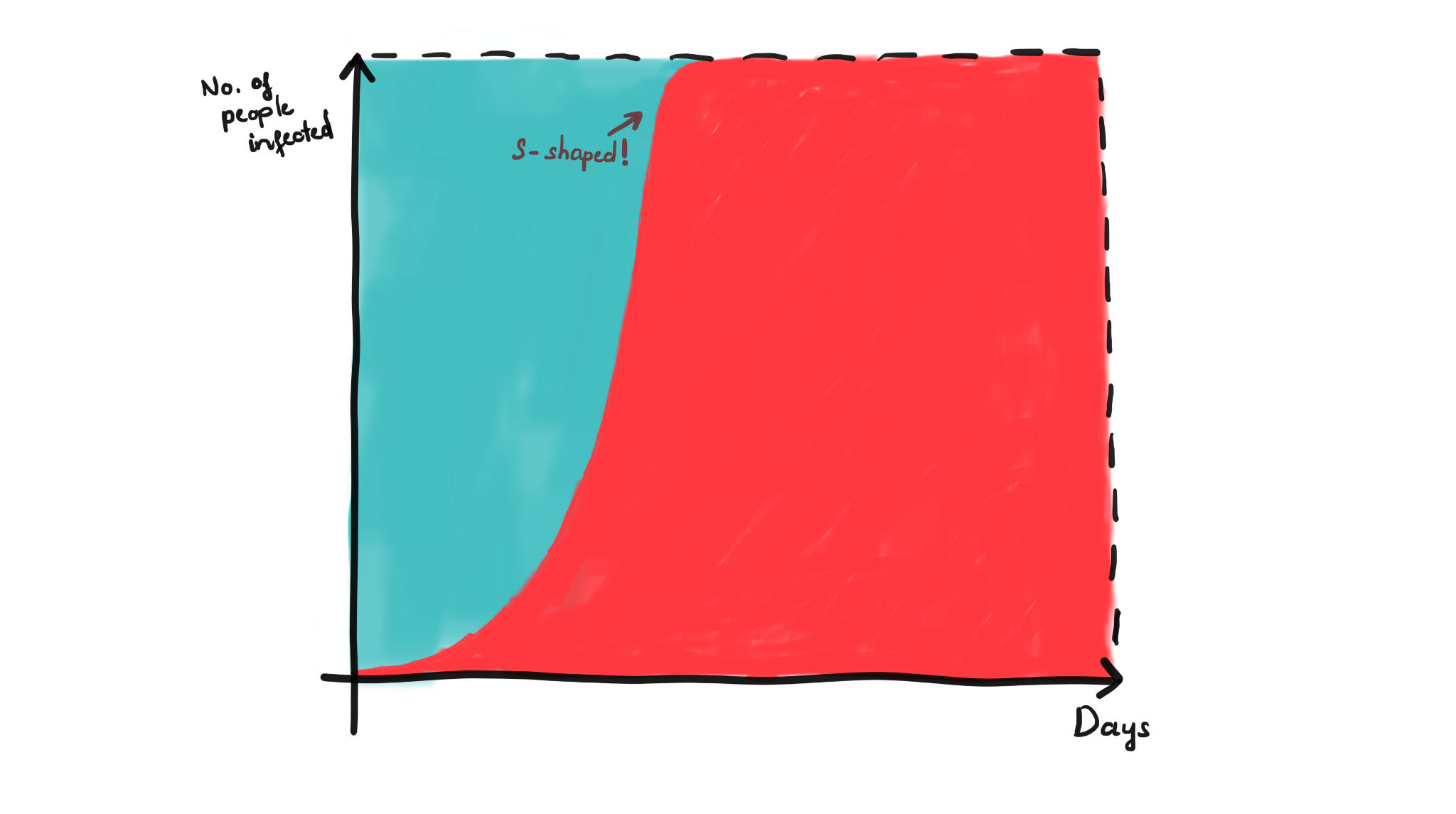
This trend follows the **basic exponential growth**, starts low then unexpectedly blows up to infinity.

Thankfully, this is definitely not happening in real life. Since the human population is fixed, the trend cannot go on forever. The idea is:



[SIR Terminology-2]

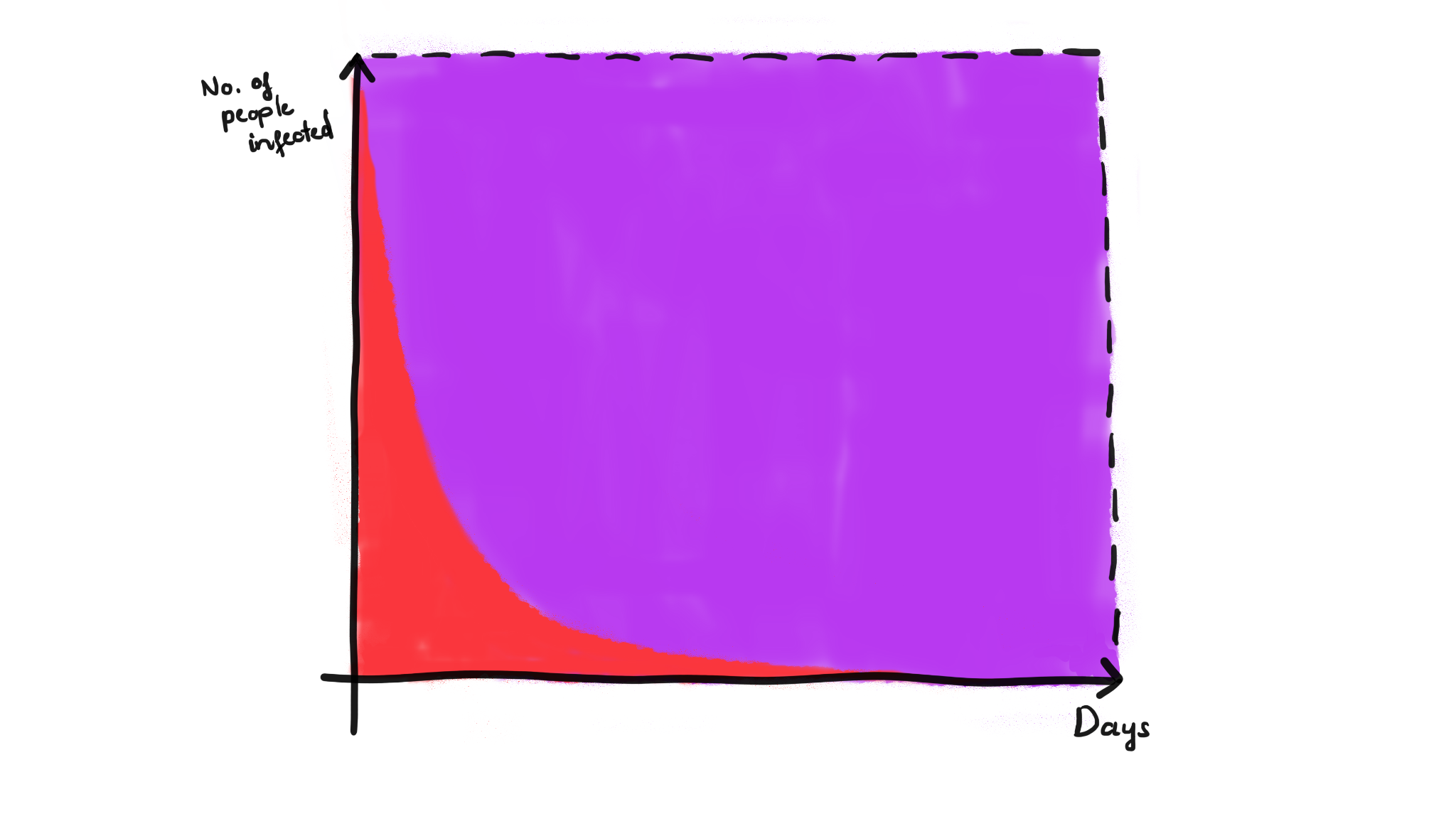
In that sense, the more **Infected**s there are, the faster **Susceptible**s become **Infected**s; but, the fewer **Susceptible**s there are, the slower **Susceptible**s become **Infected**s. The graph of epidemic growth is now changed:



[SIR Terminology-3]

This is called the **logistic growth curve**: starts low, blows up, then slows down once again. However, the concept is still not applicable. In reality, when the **Infected** stops being infectious, that means: either they are 1/ recovered (and healthy), 2/ immune to the virus but left with permanent disabilities, or in the worst case, 3/ dead. Now, for the sake of simplicity, we pretend that all **Infected** become **Recovered**, and they never be infected again.

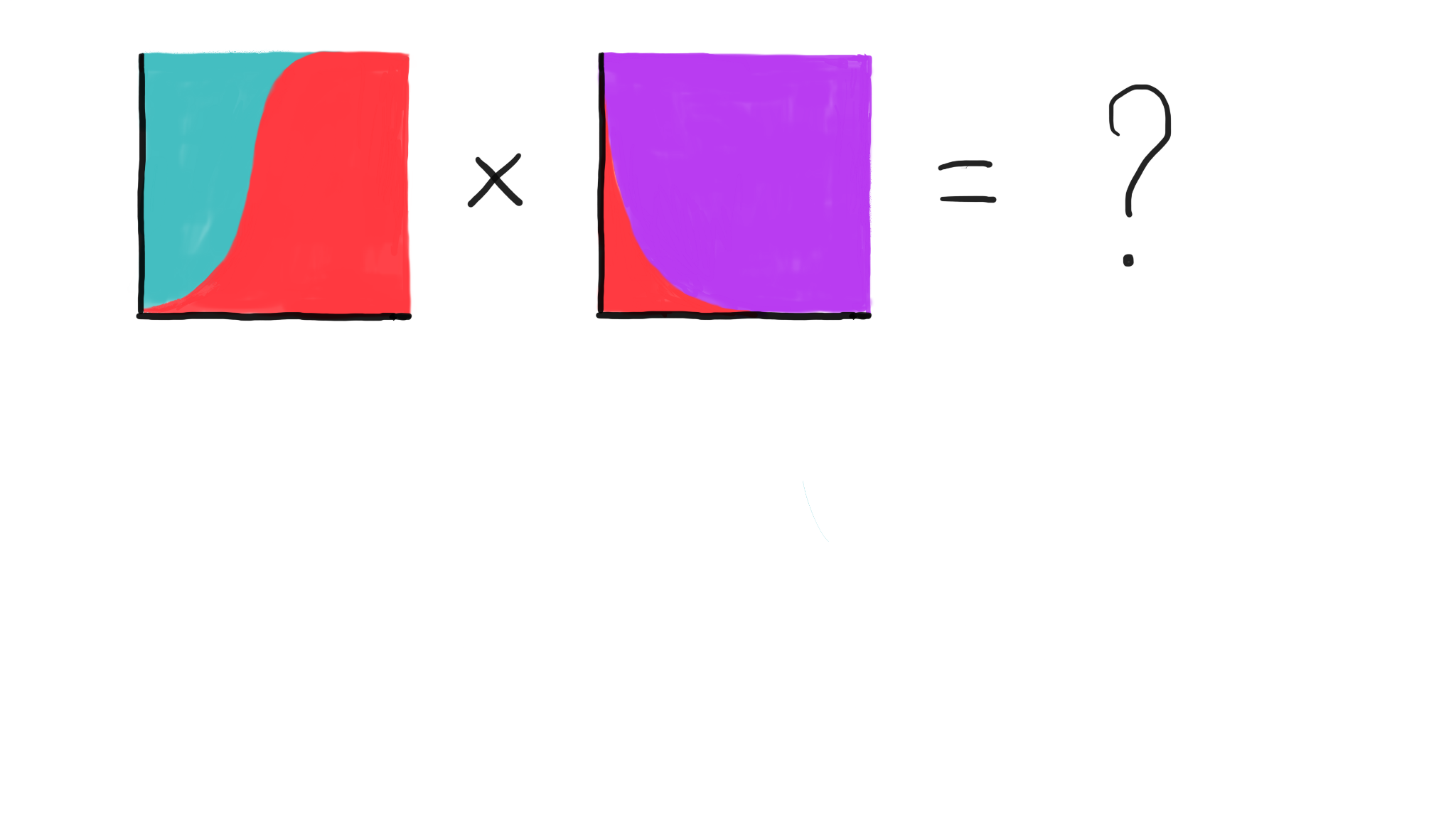
Assume that **Infected** + **Recovered** = fixed number. The change of Infected and Recovered with respect to each other is simulated below:



[SIR Terminology-4]

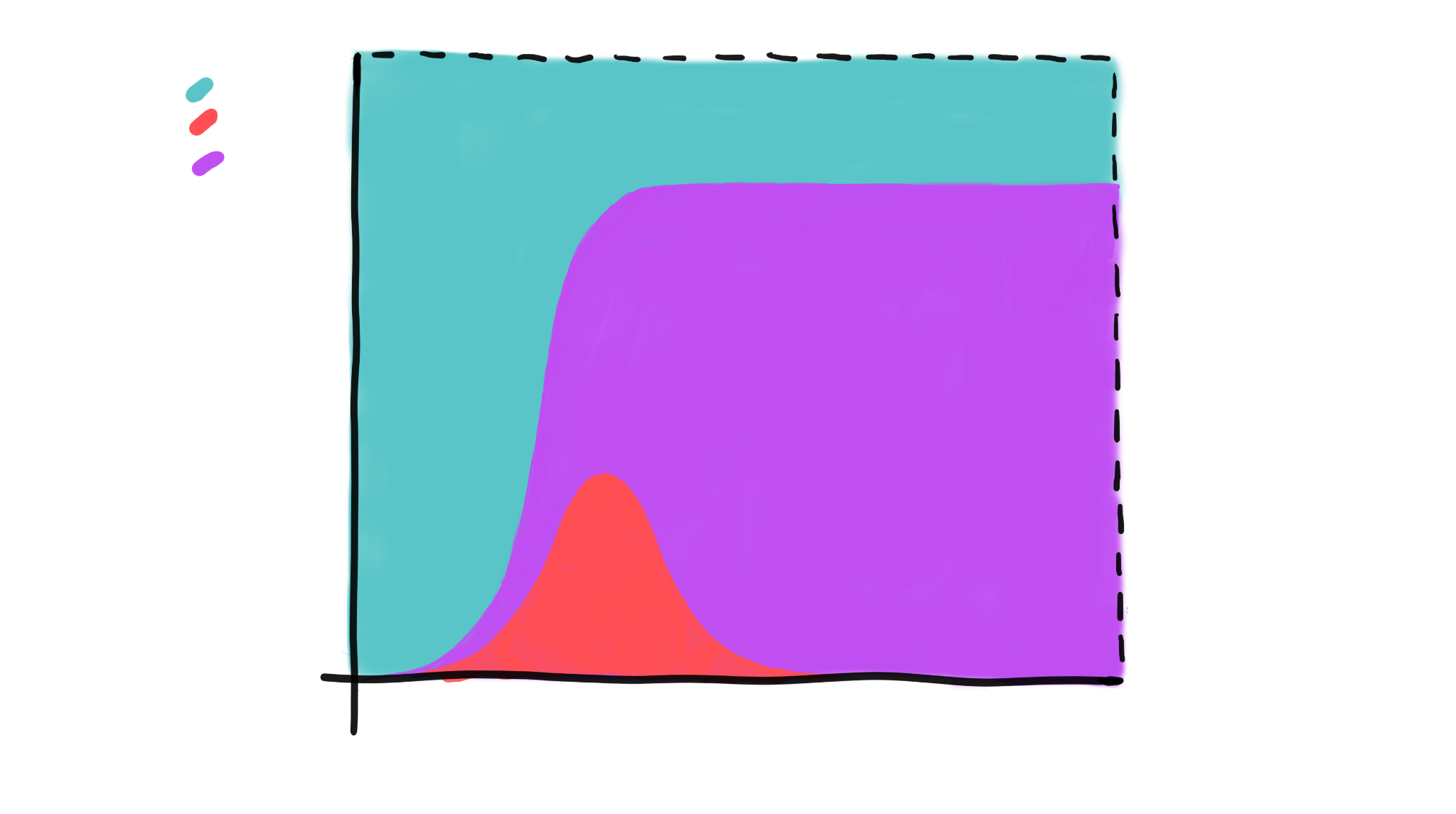
This seems opposite of the above exponential growth - the **exponential decay curve**.

Now, what if we simulate the S-shaped logistic growth (***Infected***-***Susceptible***) with decay curve (***Infected***-***Recovery***)?



[SIR Terminology-5]

Here it is:



[SIR Terminology-6]

This set of curves is actually an important and common mathematical model in data analytics - the **SIR Model** (**Susceptible** - **Infected** - **Recovered**).

1. ***SIR Model***
2. *Nguyên tắc*

SIR model divides a fixed population of N people into 3 compartments:

* Susceptible (S): The individual has not contracted the disease, but can be infected due to transmission from infected people.
* Infected (I): The individual has contracted the disease.
* Recovered (R): The individual has survived and developed immunity to the disease or is deceased.

Consider that people develop immunity and there is no transition from recovered to the remaining 2 stages, the differential equations that govern the system are:

With β is the effective contact rate of the disease and γ is the mean recovery rate. That is, each infected individual comes into contact with βN other individuals per time unit and each infected individual recovers from the disease after 1/γ time unit. (https://scipython.com/book/chapter-8-scipy/additional-examples/the-sir-epidemic-model/)

The SIR model is one of the simplest compartmental models whose main purpose is to understand the key factors that impact the epidemic transmission, which, in this case, is β and γ.

*[insert interactive video]*

1. *Observations of the model*
2. The number of susceptible individuals decreases over time.
3. The number of infected individuals reaches a peak, then decreases over time. This occurs sooner as the parameter β increases.
4. All compartments flatten out after a period of time, which occurs sooner as the parameter γ increases.
5. The number of recovered or deceased individuals increases more or less significantly as the number of infected individuals increases or decreases respectively.
6. At the start of the period, the number of infected individuals increases if β > γ and vice versa.

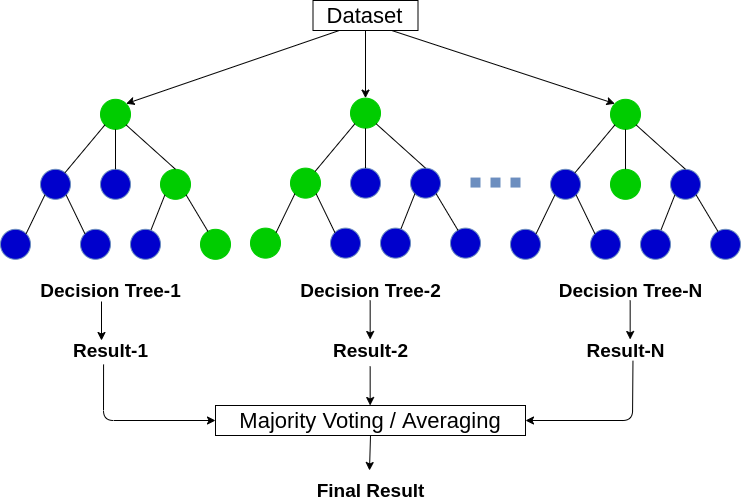
**Fifth part: Machine learning in COVID-19 prediction**

1. ***Random Forest model***
2. *Nguyên tắc*

*Because we’re trying to predict the future values, we will be using Random Forest Regression rather than Random Forest Classification*

*Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.* [*[1]*](https://levelup.gitconnected.com/random-forest-regression-209c0f354c84)

*Basically, it splits the dataset into a number of trees (n\_estimators), then it repeats the process in each tree for a number of times (max\_depth) or until all leaves are pure (all data points contain the same label). Then it takes in all of the output of the trees to form a final result.*

**

1. *14-days prediction:*

*Code + explanation:*

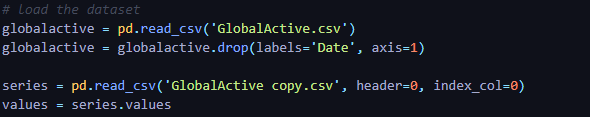
*The following code is taken from our Random Forest Regression for the Active cases on Global:*

*Random Forest Regression (RFR) requires input value and output value in order to work, so we’ll transform our time series dataset into a supervised learning dataset.*

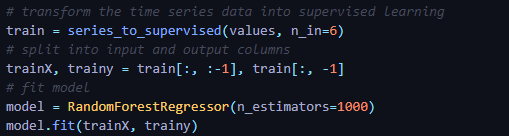
**

*(code adapted from* [*https://machinelearningmastery.com/random-forest-for-time-series-forecasting/*](https://machinelearningmastery.com/random-forest-for-time-series-forecasting/)*)*

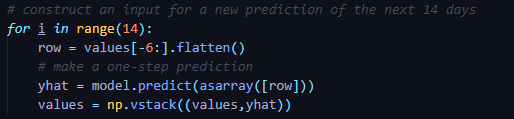
*Next, we’ll load the dataset:*

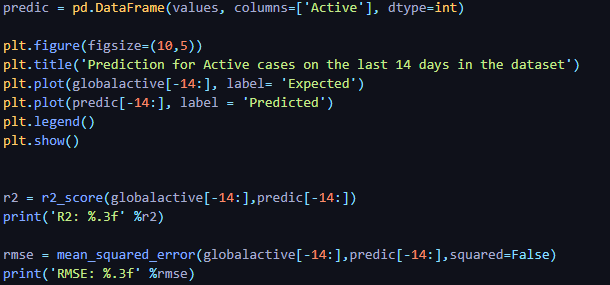
**

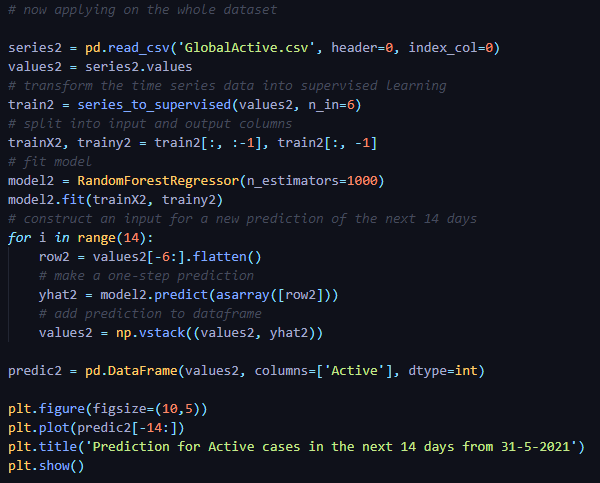
*The* ***globalactive*** *variable is only for the purpose of comparing the prediction purpose (it has the last 14 days of the dataset, until 31-5-2021) and the actual dataset used for training the model will be the* ***series*** *variable (only has data up until 17-5-2021)*

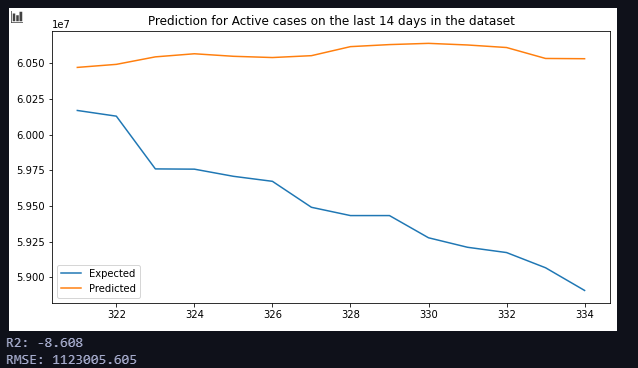
*Then, we’ll start training the model, transforming the dataset into a supervised learning one, and use the Random Forest Regressor from sklearn:*

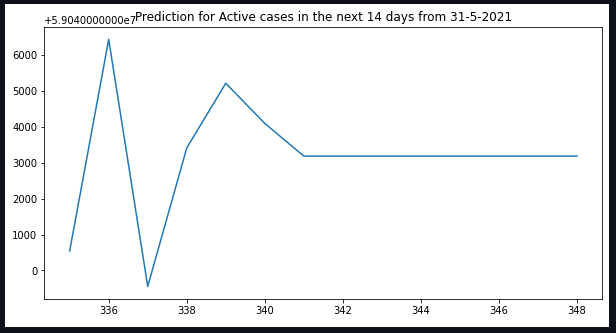
*In order to predict the value of the last 14 days, we’ll make the model predict the data of the next day from the values of the last known 6 days, then we’ll add the prediction to the dataset and do the process again, for a total of 14 times:*

**

*Then we transform the data into a dataframe and then plot it, along with the real observed data to compare, and give score on the accuracy in R2 and RMSE:*

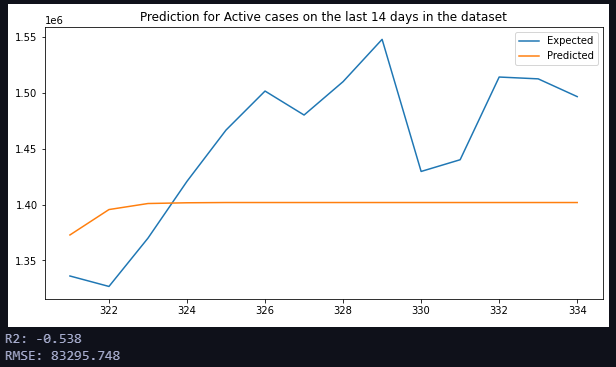
*Then we’ll apply the process again on the whole data until 31-5-2021 to see what the values is predicted to be after that, which is our main purpose:*

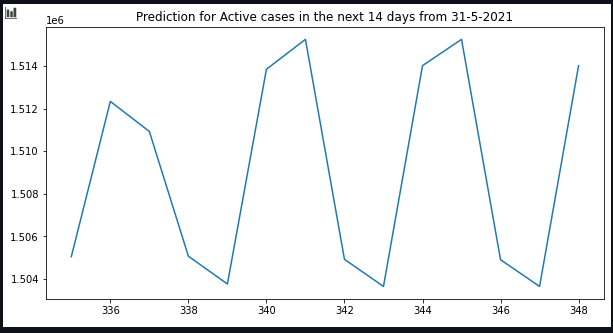
*After running the code, we get the evaluation:*

*And the prediction for what the value will be 14 days after 31-5-2021:*

*Brazil:*

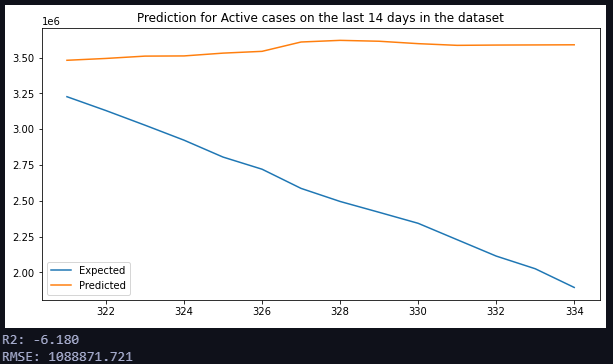
*The predicted value changes the most on the first day, then remains stable. As we can see, the predicted value is far from the expected value, so the scores are not looking very good.*

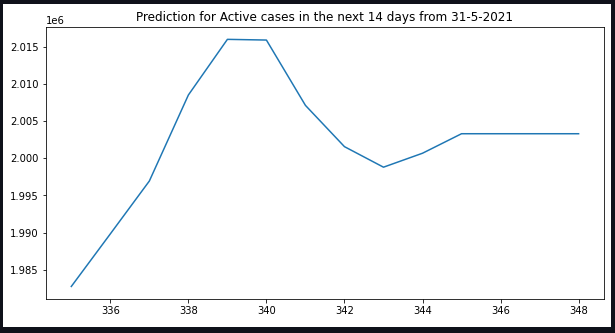
**

*And for this prediction, we can see that the predicted data is in a stationary form, which is what we know is not possible for a pandemic.*

*India:*

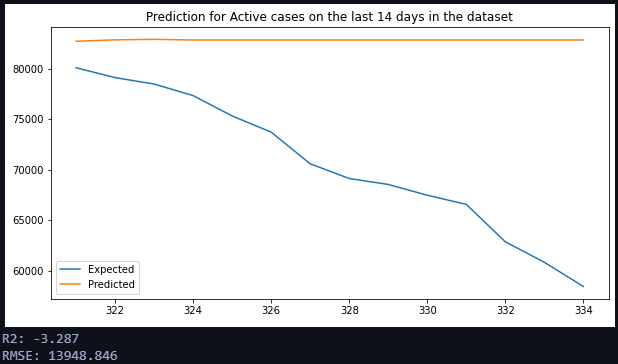
*The predicted value is going slightly upward, while the expected value is going downward which is the opposite trend.*

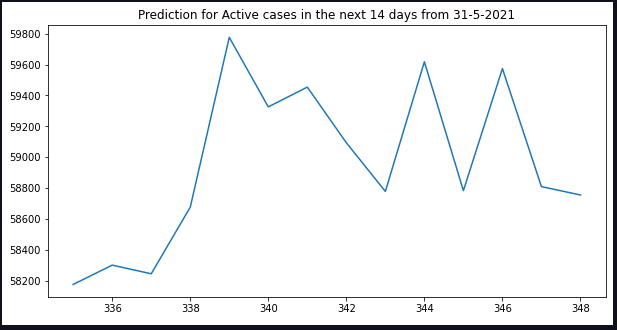
**

*The prediction for future value increases on the first few days, then decreases and goes up again. It becomes stable on the last days, which is virtually impossible with pandemics.*

*Japan:*

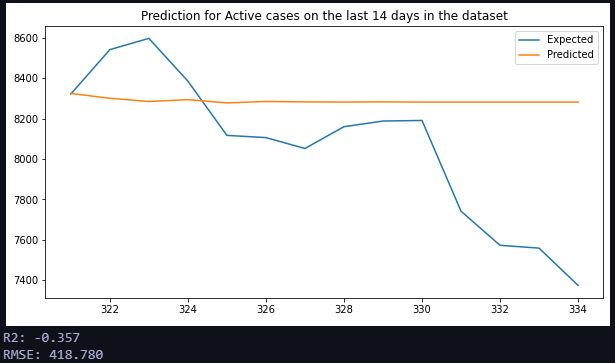
*The predicted value does change overtime, although it is almost unnoticeable. It is still far from the expected value, so the scores are not very positive.*

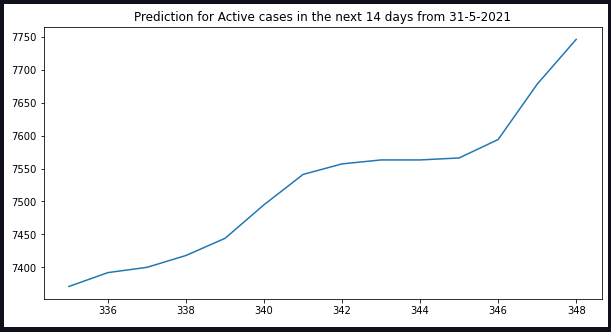
**

*The prediction for the next 14 days changes drastically each day and does not seem to follow any pattern, which is probable for a pandemic, but still not very convincing.*

*South Korea:*

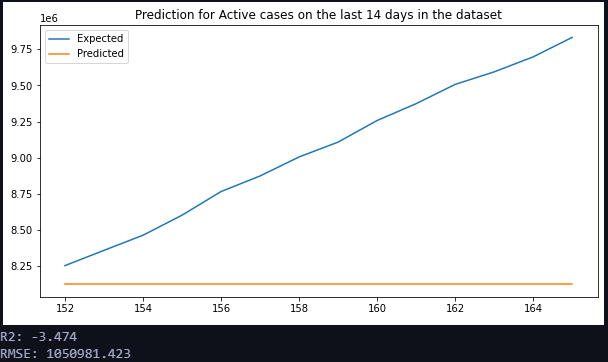
*Although the prediction for the first day is very close, the predicted value does not experience any notable changes over time while the expected value is shifting a lot.*

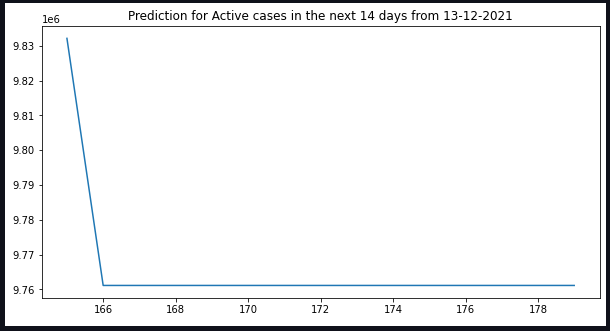
**

*This prediction is looking very convincing for a pandemic: an upward trend. However, the change in number between days is still very minuscule for a big pandemic like Covid-19*

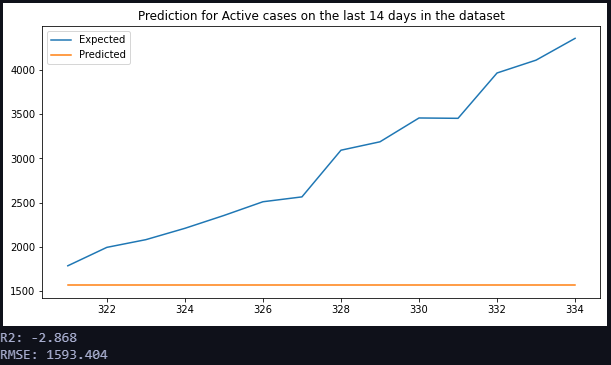
*US:*

*The predicted value remains stable while the expected value is increasing steadily, so the accuracy is not looking very optimistic.*

**

*The predicted value has a sharp decrease on the first day then remains stable for the rest of the time frame. This does not look very likely for a pandemic like Covid-19.*

*Vietnam:*

**

**

*\_ Observations (trend gì? Tăng? Giảm? Nhiều ko?)*

*In the evaluation, the predicted value is seemingly going in an opposite trend as the expected value.*

*In the prediction for the next 14 days after 31-5-2021, the value fluctuates around 3000 more cases then becomes stable.*

1. ***ARIMA Model***
2. *Definition*

*\_ Nguyên tắc hoạt động*

* ARIMA, short for ‘Auto Regressive Integrated Moving Average’ is actually a class of models that ‘explains’ a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

Cite: https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/

*\_ Các khái niệm p, d, q*

* ‘p’ is the order of the ‘Auto Regressive’ (AR) term. It refers to the number of lags of Y to be used as predictors
* The value of d, therefore, is the minimum number of differencing needed to make the series stationary.
* ‘q’ is the order of the ‘Moving Average’ (MA) term. It refers to the number of lagged forecast errors that should go into the ARIMA Model.

Cite: https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/

*\_ Autocorrelation gram (which point did you use to lag?)*

* Autocorrelation refers to the degree of correlation of the same variables between two successive time intervals. It measures how the lagged version of the value of a variable is related to the original version of it in a time series.
* Autocorrelation, as a statistical concept, is also known as serial correlation. It is often used with the autoregressive-moving-average model (ARMA) and autoregressive-integrated-moving-average model (ARIMA). The analysis of autocorrelation helps to find repeating periodic patterns, which can be used as a tool of technical analysis in the [capital markets](https://corporatefinanceinstitute.com/resources/knowledge/trading-investing/capital-markets/).
* In many cases, the value of a variable at a point in time is related to the value of it at a previous point in time. Autocorrelation analysis measures the relationship of the observations between the different points in time, and thus seeks for a pattern or trend over the time series. For example, the temperatures on different days in a month are autocorrelated.
* Similar to [correlation](https://corporatefinanceinstitute.com/resources/knowledge/finance/correlation/), autocorrelation can be either positive or negative. It ranges from -1 (perfectly negative autocorrelation) to 1 (perfectly positive autocorrelation). Positive autocorrelation means that the increase observed in a time interval leads to a proportionate increase in the lagged time interval.

Cite: https://corporatefinanceinstitute.com/resources/knowledge/other/autocorrelation/

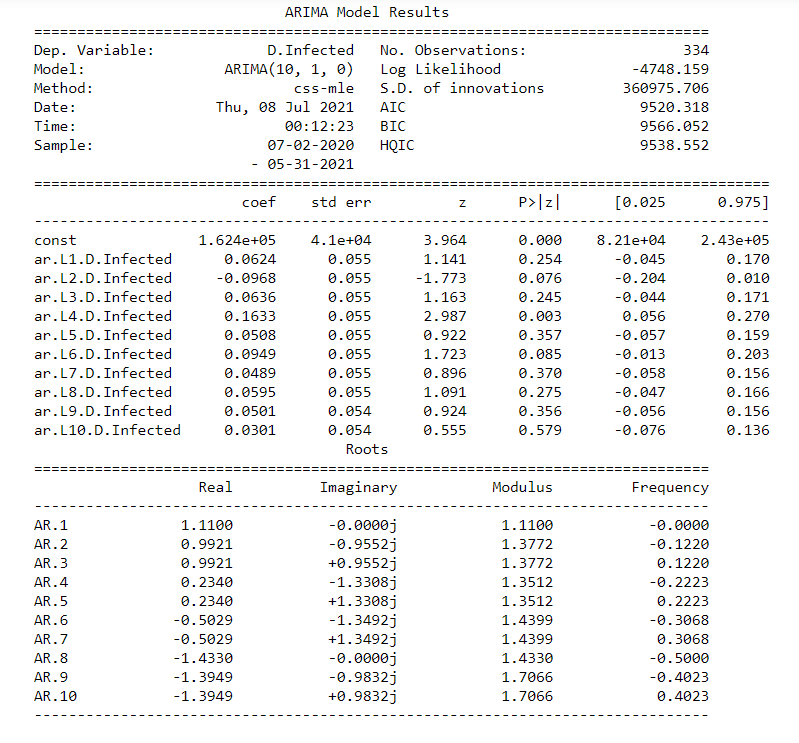
1. *K/q training*

*\_ Code + giải thích + chức năng (theo cell)*

Order (10,1,0)

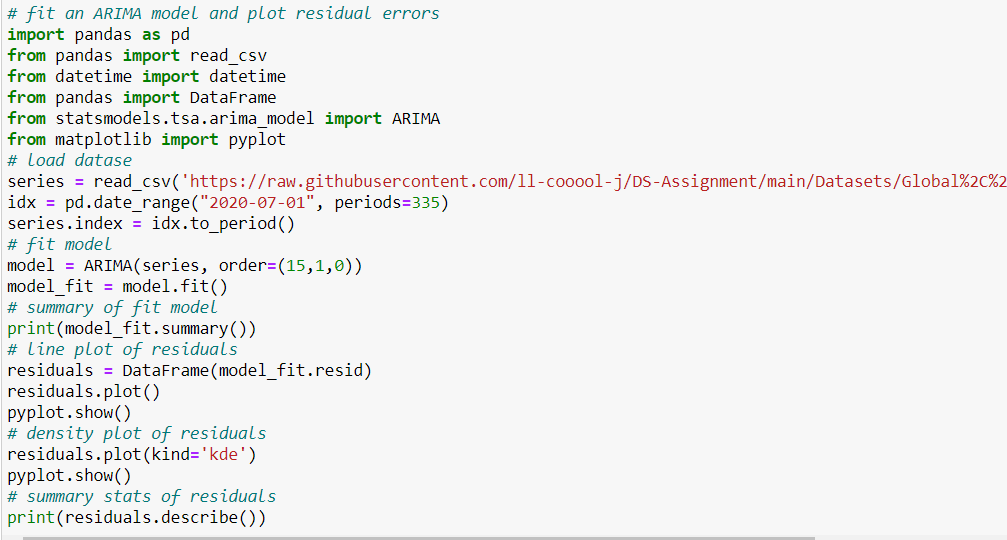
**

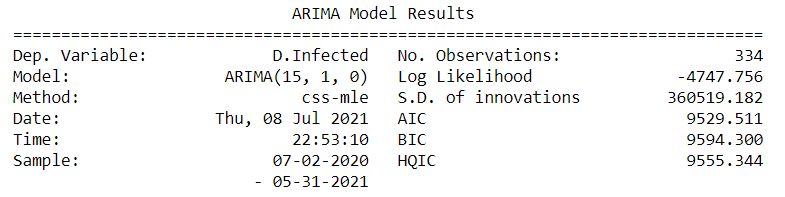
*\_ Xuất ra toàn bộ bảng + hình trong code (khỏi giải thích đi huhu T\_T)*

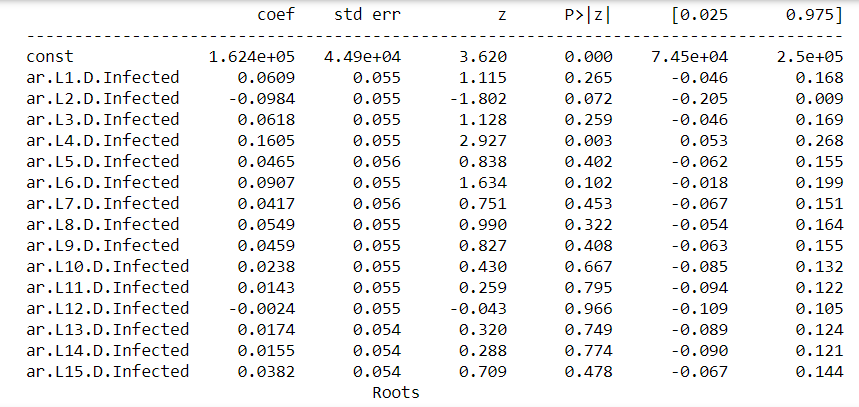
**

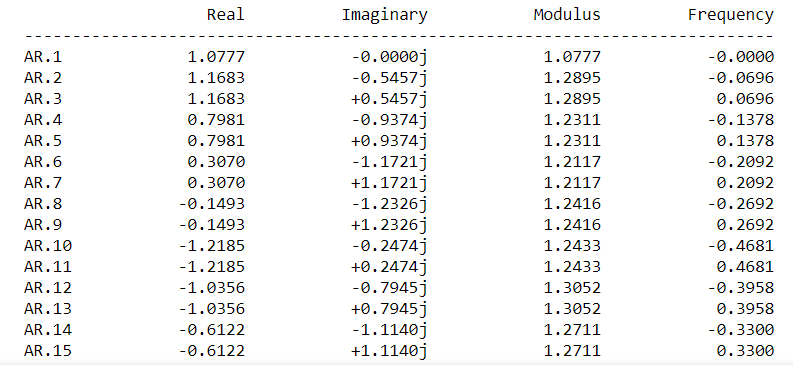
**

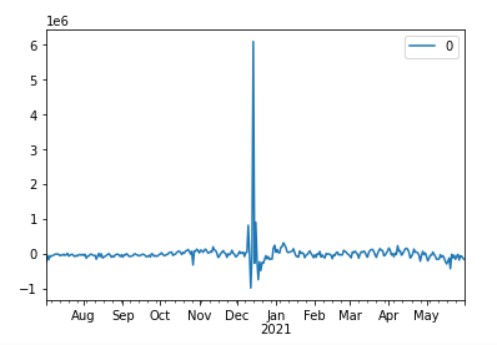
Order(15,1,0)











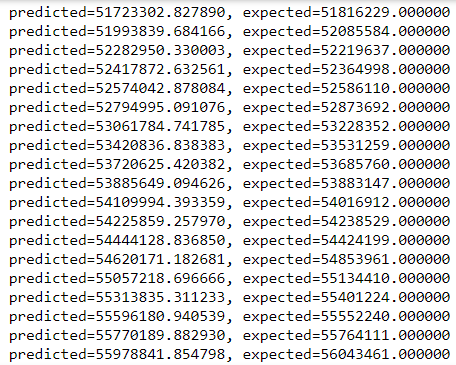


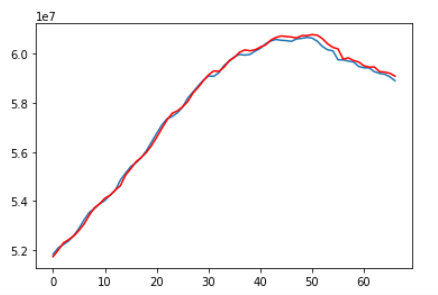
1. *K/q predict*

*\_ Code + giải thích + chức năng (theo cell)*

**

*\_ Bảng prediction + plot*

**

**

1. ***RapidMiner Automodel***

* *Sơ lược các terms & thao tác trong tool*
* *Các kq model*

1. ***Other models that we didn't actually take in***

* *FB Prophet*
* *C4.5 Algorithm*
* *Other SIR variations*

**Sixth part: Discussion**

1. ***Version control records***
2. *Ver 1 of project (đã làm gì?) - ver 1 là trước meeting ngày 3/7*

*\_ SIR ver 1: Code + giải thích (theo cell) & code plot đã làm*

Code adapted from [Scipython The SIR epidemic model](https://scipython.com/book/chapter-8-scipy/additional-examples/the-sir-epidemic-model/)

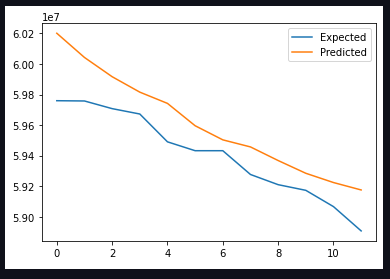
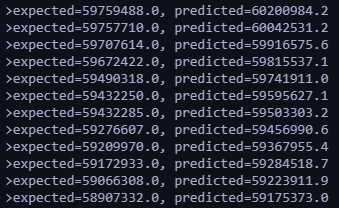
[SIR ver 1 ~ a.png] [SIR ver 1 ~ b.png]

*\_ Random Forest: m*

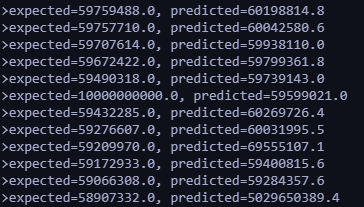
*In version 1, we copied the whole code from* [*Random Forest for Time Series Forecasting (machinelearningmastery.com)*](https://machinelearningmastery.com/random-forest-for-time-series-forecasting/)*.*

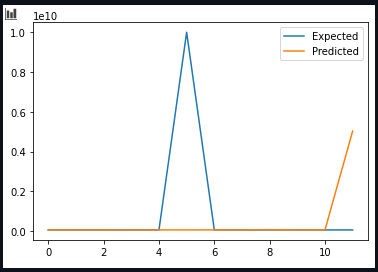
*However, after running and testing, we found out that it doesn’t quite fit what we’re trying to do.*

*The code used walk-forward-validation, which means that after it gives a prediction, it will take in the real value (instead of the predicted value) to use for the next prediction.*

*For example, below is the predicted value for the last 12 days of the dataset:*

*However, if we change the value of the dataset of the last 12 days, it will affect the next prediction:*

**

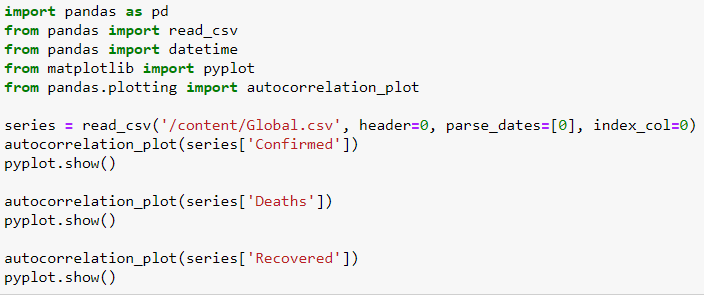
**

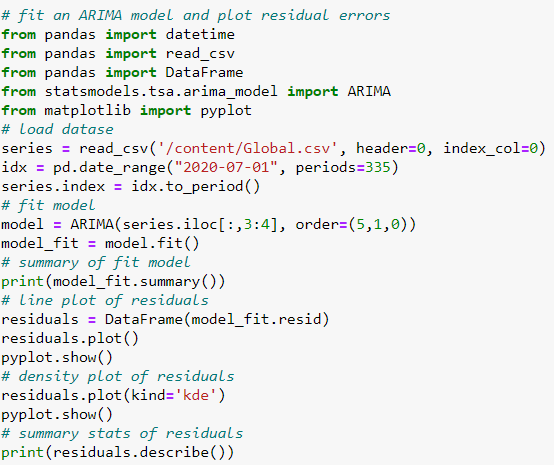
*We changed the value on the sixth day, and the value of the 12th day became abnormally high. Therefore, the real value actually affects the prediction, so it doesn’t fit what we were aiming for.*

*So we changed our approach. We only keep transforming a dataset into a supervised learning one algorithm because we can use it for our new version. We then made a new code based on what we guessed would suit: using the data of the last 6 days to give a prediction and then add that prediction to the data and do it again.*

*\_ ARIMA: Code + giải thích (theo cell) & show các lỗi*

**

**

**

**

1. *Ver 2 of project*

*\_ SIR ver 2: Note các thay đổi + giải thích*

Split code to plot 2 models susceptible and infected/recovered due to the possible difference between 3 compartments.

[SIR ver 2 ~ a.png] [SIR ver 2 ~ b.png]

Apply code on 3 periods

*\_ RDFR + ARIMA: smol changes*

1. *Ver 3 of project*

*\_ New observations (phát hiện ra mình hiểu sai chức năng của SIR) -> sử dụng p/a thay thế là Geogebra*

Change from python code to geogebra interactive model

*\_ RDFR + ARIMA: ko đổi, như trên*

1. ***Project limitations***
2. *SIR:*

*\_ Beta và gamma có cách tính rất phức tạp -> cần nghiên cứu riêng và sâu hơn mới observe được*

*\_ R0 cũng z*

*\_ Limited usage of SIR model: chỉ dùng để áng chừng tương lai, ko dự đoán chính xác dc, hơn nữa beta và gamma cũng nhảy xoành xoạch. Tuy nhiên nó vẫn có ích cho các nhà chính sách và chính phủ trong việc đưa ra quyết định.*

The SIR model is used to estimate the spreading of the epidemic and provides a less accurate result compared to other predictive models due to the fact that β and γ are time dependent. Furthermore, we have to go through a complicated process (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7570398/) to determine β and γ so that the result is closest to reality, so this seems not very practical for reality forecasting purposes.

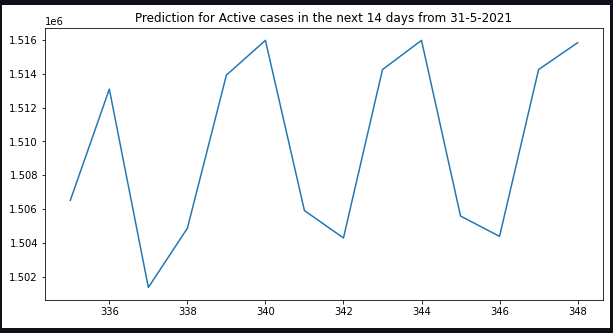
However, the SIR model is still one of the most important and significant mathematical models in the [first steps] of dealing with pandemics. It is used to see how effective the current response to an epidemic is, thus suggesting a better policy.

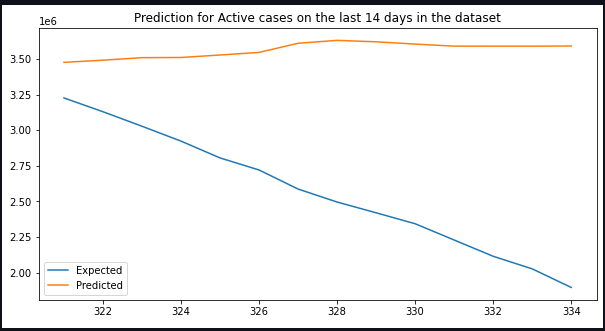
1. *Random Forest:*

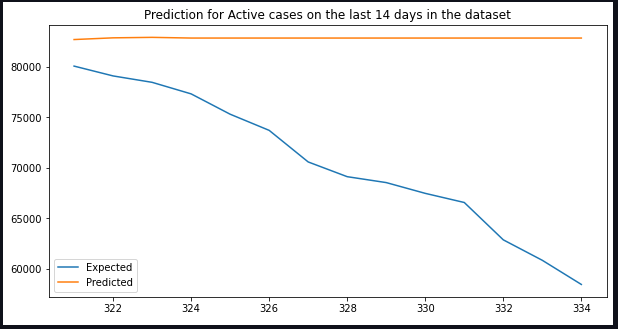
*Due to its characteristics, Random Forest is good with Stationary datasets, like Average temperature of a house, but not good with Non-stationary datasets, like this Covid-19. Therefore, the results we got are overfitted.*

*(Overfitting is when the model is following its training data too closely, which prevent the model from giving any reliable predictions because the predicted value becomes stationary or fluctuate around a certain point)*

*Below is the prediction for the Active cases in Brazil. As we can see, the result is in a Stationary form while what we are expecting is in a Non-stationary form.*

**

*This is our prediction for evaluating in India. The real value is going downward while the predicted value is going slightly upward. (Opposite trend)*

*This is our prediction for evaluating in Japan. The real value is going down a lot while the predicted value is only changing unnoticeably. (Doesn’t change much)*

*Also, because of vague definitions of "recovered" and the fact that some states do not give data correctly, the provider of our dataset decided to stop getting the data on the recovered cases of the US from 13-12-2020. Therefore, we can only use data up until 13-12-2020.*

1. *ARIMA*

*\_ Code còn lỗi (kèm ảnh)*

*\_ Giải thích về autocorrelation gram là cảm tính chứ chưa biết đúng hay không*

*\_ Chưa rõ ràng về p d và q*

1. *What could have been done?*

\_ Có một số nguồn tài liệu dc tìm đến khá trễ -> việc làm quen các khái niệm và path làm việc của một dự án ntn diễn ra chậm hơn do với tiến độ

\_ Tìm hiểu cách dùng Github hơi muộn -> việc trao đổi file giữa nhóm ko thuận lợi

\_ Thời gian tìm hiểu các model khá lâu -> ko chọn lọc đc các model đã được confirm là có accuracy cao hơn

1. ***Team captain’s notes***

\_ Log sử dụng công cụ hỗ trợ manage team

\_ Meeting records (note là có đính kèm mấy file slides)

**Seventh part: Bottom sections that every report has but nobody ever reads and never will**

1. ***Acknowledgement (done)***
2. ***References***

Source code SIR : https://scipython.com/book/chapter-8-scipy/additional-examples/the-sir-epidemic-model/