Title: COVID-19 Progression Prediction.

Objective: The outbreak of Covid-19 is developing into a major international crisis, and it's starting to influence important aspects of daily life. A strong model that predicts how the virus could spread across different countries and regions may be able to help mitigation efforts. The goal of this task is to build a model that predicts the progression of the virus throughout March 2020.

Tools and Dataset: The source of data: KCDC (Korea Centers for Disease Control & Prevention). This dataset contains 7k+ patient information with 15 different features.

- id: Unique id of the patient.
- sex: Sex/Gender of the patient.
- birth_year: Birth year of the patient.
- country: Country of the patient.
- region: Residential region of the patient.
- Disease: 0: no disease / 1: underlying disease
- group: The collective infection.
- infection_reason: How the patient got infected.
- infection_order: The order of infection.
- infected_by: The ID of the patient who infected this patient.
- contact_number: The number of contacts with people.
- confirmed_date: The date of confirmation that people is infected.
- released_date: The date of discharge.
- deceased_date: The date of decease.
- state: The current state of the patient (isolated, released and deceased).

Furthermore, Jupyter Notebook (5.4.0) used for the Python code execution and sklearn libraries used to implement regression.

Description of the procedure:

- The number of cases from one day to next day are completely random as the number of cases increases day by day are independent of each other.
- As of now let's assume number of new cases each day is proportional to the number of existing cases, it means each day it's get multiplied by a constant.
- Intuitively it means as the date changes, the number of confirmed cases also increases as they both are directly proportional to each other.
- So, if we compare total cases from one day to next day, then tracking the changes between number of cases is nothing but the growth factor.
- Simply growth factor is the ratio between two successive changes and that resultant ratio is the constant that get multiplied each day.
- So, with existing accumulated data(number of cases each day), we'll predict the expected number of cases for future dates by using Linear Regresion which is one of the simplest but powerful concept of machine learning.

  (**Note:** Coding and Figures are represented in Notebook file.)

Result and discussion:

- The given dataset is collected by KCDC for Korea, China and Mongolia country.
- Only 0.4% of patient died but also only 0.7% of patient recovered and still around 98.9% of patient are under isolation.
- So, the death probability is low but also recovering from this virus is difficult.
- The most of the patient died within a 5 days after confirmation. So, the treatment of corona virus has to be started immediately after confirmation as it's impact is really hazardous.
- The patient with age between 35 and 45 years is more likely to get released but this is not true in all cases.
- So, even though death percentage is low but recovering from this virus is difficult.