Open in Colab
(https://colab.research.google.com/github/Elzawawy/covid-case-estimator/blob/master/Country_Data_Exploration.ipynb)

COVID-19 Country Data Exploration

- Under the light of learning more about COVID-19 and how different countries are affected by it and why. It might be useful to compare different metrics between different countries.
- In this dataset we dig very deep into more complicated and diverse per-country features in a trial to improve Model Class 2 and develop more and better features for it.
- This dataset is thanks to @Patrick from Kaggle. You can find it here (https://www.kaggle.com/bitsnpieces/covid19-country-data/version/12?select=covid19_data).

In [1]:

```
#imports cell
import pandas as pd
import numpy as np
import pickle
from shutil import copyfile

# Plotting libraries
import matplotlib.pyplot as plt
import plotly.express as px

# mount google drive to copy files from repo into drive.
from google.colab import drive
drive.mount('/content/drive')
STORAGE_DIR = "/content/drive/My Drive/COVID-19/country-data/"
```

```
Go to this URL in a browser: https://accounts.google.com/o/oauth2/auth?client_id=947318989803-6
bn6qk8qdgf4n4g3pfee6491hc0brc4i.apps.googleusercontent.com&redirect_uri=urn%3aietf%3awg%3aoauth
%3a2.0%3aoob&response_type=code&scope=email%20https%3a%2f%2fwww.googleapis.com%2fauth%2fdocs.te
st%20https%3a%2f%2fwww.googleapis.com%2fauth%2fdrive%20https%3a%2f%2fwww.googleapis.com%2fauth%
2fdrive.photos.readonly%20https%3a%2f%2fwww.googleapis.com%2fauth%2fpeopleapi.readonly

Enter your authorization code:
··········
Mounted at /content/drive
```

Downloading Dataset

- We use the Official API for https://www.kaggle.com (https://www.kaggle.com) to get our datasets.
- You can get your own Kaggle API key to run this cell by going to kaggle.com and navigating to `My Account` Tab and use the `Create API Key` button, you then upload it to the notebook's temproray storage.

In [3]:

```
!pip install kaggle
# You have to upload you own Kaggle API which is the `kaggle.json` into the temp directory first.
!cp /content/kaggle.json ~/.kaggle/kaggle.json
# For the Kaggle API key to be un-readable by other users on this system.
!chmod 600 /root/.kaggle/kaggle.json
!kaggle datasets download -d bitsnpieces/covid19-country-data
!mkdir country_data_dataset
!unzip covid19-country-data.zip -d country_data_dataset
!rm covid19-country-data.zip
```

```
Requirement already satisfied: kaggle in /usr/local/lib/python3.6/dist-packages (1.5.6)
Requirement already satisfied: urllib3<1.25,>=1.21.1 in /usr/local/lib/python3.6/dist-packages
(from kaggle) (1.24.3)
Requirement already satisfied: python-slugify in /usr/local/lib/python3.6/dist-packages (from k
aggle) (4.0.0)
Requirement already satisfied: python-dateutil in /usr/local/lib/python3.6/dist-packages (from
kaggle) (2.8.1)
Requirement already satisfied: certifi in /usr/local/lib/python3.6/dist-packages (from kaggle)
(2020.4.5.1)
Requirement already satisfied: tqdm in /usr/local/lib/python3.6/dist-packages (from kaggle) (4.
41.1)
Requirement already satisfied: six>=1.10 in /usr/local/lib/python3.6/dist-packages (from kaggle
) (1.12.0)
Requirement already satisfied: requests in /usr/local/lib/python3.6/dist-packages (from kaggle)
(2.23.0)
Requirement already satisfied: text-unidecode>=1.3 in /usr/local/lib/python3.6/dist-packages (f
rom python-slugify->kaggle) (1.3)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.6/dist-packages (fro
m requests->kaggle) (3.0.4)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.6/dist-packages (from req
uests->kaggle) (2.9)
Downloading covid19-country-data.zip to /content
  0% 0.00/186k [00:00<?, ?B/s]
100% 186k/186k [00:00<00:00, 27.6MB/s]
Archive:  covid19-country-data.zip
  inflating: country_data_dataset/country_names_covid19_forecast.csv
  inflating: country_data_dataset/covid19_data - 2009_flu_pandemic.csv
  inflating: country_data_dataset/covid19_data - age.csv
  inflating: country_data_dataset/covid19_data - airport_traffic.csv
  inflating: country_data_dataset/covid19_data - airport_traffic_world.csv
  inflating: country_data_dataset/covid19_data - cost_of_living.csv
  inflating: country_data_dataset/covid19_data - covid19_strains.csv
  inflating: country_data_dataset/covid19_data - covid_tests.csv
  inflating: country_data_dataset/covid19_data - data_sources.csv
  inflating: country_data_dataset/covid19_data - flu_pneumonia_death.csv
  inflating: country_data_dataset/covid19_data - gdp.csv
  inflating: country_data_dataset/covid19_data - health.csv
  inflating: country_data_dataset/covid19_data - hospital_beds.csv
  inflating: country_data_dataset/covid19_data - lat_long.csv
  inflating: country_data_dataset/covid19_data - population.csv
  inflating: country_data_dataset/covid19_data - property_prices.csv
  inflating: country_data_dataset/covid19_data - quality_of_life.csv
  inflating: country_data_dataset/covid19_data - school_closures.csv
  inflating: country_data_dataset/covid19_data - sex.csv
  inflating: country_data_dataset/covid19_merged.csv
  inflating: country_data_dataset/dhl_people_breadth.csv
```

Understanding the Dataset

- The dataset has a lot of features to try out and check the correlation between them and the `total_cases` .
- Features avaiable from the dataset:
  - Monthly temperature and precipitation from Worldbank.
  - Latitude and longitude
  - Population, density, gender and age
  - Airport traffic from Worldbank
  - COVID-19 date of first case and number of cases and deaths as of March 26, 2020
  - 2009 H1N1 flu pandemic cases and deaths obtained from Wikipedia
  - Property affordability index and Health care index from Numbeo
  - Number of hospital beds and ICU beds from Wikipedia
  - Flu and pneumonia death rate from Worldlifeexpectancy.com (Age Adjusted Death Rate Estimates: 2017)
  - School closures due to COVID-19
  - Number of COVID-19 tests done
  - Number of COVID-19 genetic strains
  - US Social Distancing Policies from COVID19StatePolicy's SocialDistancing repository on GitHub
  - DHL Global Connectedness Index 2018 (People Breadth scores)

In [0]:

```
### Load Total Cases to check correlation with features.
TOTAL_CASES_FILE_PATH = "/content/drive/My Drive/COVID-19/our-world-in-data/country-features/total_cases_dict.pickle"
# Load data (deserialize) from pickle file.
with open(TOTAL_CASES_FILE_PATH, 'rb') as handle:
    total_cases_dictionary = pickle.load(handle)

# a safe getter fucntion for lists in Python 3.xx
def safe_list_get (l, idx, default):
  try:
    return l[idx]
  except IndexError:
    return default

# matcher function that matches a list of keys with the total_cases_dict keys to join them.
def matcher(k):
    x = (i for i in total_cases_dictionary if i ==k)
    return safe_list_get(list(map(total_cases_dictionary.get, x)),0,np.nan)

# helper method to save dicts into pickle files.
def save_dict_to_pickle(dict, pickle_file):
  with open(pickle_file, 'wb') as handle:
    pickle.dump(dict, handle, protocol=pickle.HIGHEST_PROTOCOL)
```

Trying out Country GDP Features

- **Gross Domestic Product (GDP)** is the monetary value of all finished goods and services made within a country during a specific period. GDP provides an economic snapshot of a country, used to estimate the size of an economy and growth rate. GDP can be calculated in three ways, using expenditures, production, or incomes.
- We will see how this **feature correlates with the number of total cases for each country.**

In [6]:

```
GDP_FEATURE_FILE_PATH = "/content/country_data_dataset/covid19_data - gdp.csv"
gdp_dataframe = pd.read_csv(GDP_FEATURE_FILE_PATH)
gdp_dataframe.head()
```

Out[6]:

	Rank	Country/Territory	GDP_USD_Million
0	1	United States	21,439,453
1	2	European Union	18,705,132
2	2	China	14,140,163
3	3	Japan	5,154,475
4	4	Germany	3,863,344

Adding Total Cases Column to Dataframe

In [11]:

```
gdp_dataframe['total_cases'] = gdp_dataframe['Country/Territory'].map(matcher)
gdp_dataframe['GDP_USD_Million'] = gdp_dataframe['GDP_USD_Million'].str.replace(',','').astype(float)
gdp_dataframe.dropna(inplace=True)
gdp_dataframe.head()
```

Out[11]:

	Rank	Country/Territory	GDP_USD_Million	total_cases
0	1	United States	21439453.0	1467884
2	2	China	14140163.0	84044
3	3	Japan	5154475.0	16285
4	4	Germany	3863344.0	174355
5	5	India	2935570.0	90927

Correlation Between GDP and Total Cases

- Plotting Scatter Plot.
- Calculate correlation factor. (Value = 0.832602)

Observations:

- There's a high correlation between them and GDP sounds as an effective feature to add for country features.

In [14]:

```
px.scatter(gdp_dataframe,'GDP_USD_Million','total_cases',title = "Total Cases vs GDP").show()
gdp_dataframe[gdp_dataframe.columns[:]].corr()['total_cases'][:]
```

Out[14]:

```
Rank              -0.315420
GDP_USD_Million    0.832602
total_cases        1.000000
Name: total_cases, dtype: float64
```

Saving GDP Feature File

In [0]:

```
countries = gdp_dataframe['Country/Territory'].unique()
country_gdp_dict = {}
for country in countries:
  country_gdp_dict[country] = gdp_dataframe[gdp_dataframe['Country/Territory'] == country]['GDP_USD_Million'
].to_numpy()[0]
# save file to permenant storage on drive.
save_dict_to_pickle(country_gdp_dict, STORAGE_DIR+"gdp_dict.pickle")
```

Trying out 2009 H1N1 Features.

- In 2009, a simialr pandemic called `H1N1 flu pandemic` has occured.
- There are **number of cases and deaths** obtained from Wikipedia.
- We will check **if there is a correlation** between this old virus that envaded the planet and Today's virus envading it as well.

In [20]:

```
H1N1_FILE_PATH = "/content/country_data_dataset/covid19_data - 2009_flu_pandemic.csv"
h1n1_dataframe = pd.read_csv(H1N1_FILE_PATH)
h1n1_dataframe.head()
```

Out[20]:

	Country	Geographic_spread	Intensity	Impact_on_healthcare_services	Cases_underestimate	Cases_confirmed_
0	United States	W	**	mod	1	1
1	Brazil	R	*	mod	1	
2	India	W	*	low	0	
3	Mexico	W	**	mod	0	
4	China	NaN	NaN	NaN	0	1

Adding Total Cases Column to Dataframe.

- Additionally, Dropping useless columns and converting string columns to floats.

In [22]:

```
# add new total cases column.
h1n1_dataframe['total_cases'] = h1n1_dataframe['Country'].map(matcher)
# drop useless columns.
h1n1_dataframe.drop(columns=['Geographic_spread',
                             'Intensity',
                             'Impact_on_healthcare_services',
                             'Cases_underestimate'], inplace=True)
# convert string cols to floats.
h1n1_dataframe['Cases_confirmed_clean'] = h1n1_dataframe['Cases_confirmed_clean'].str.replace(',','').astype
(float)
h1n1_dataframe['Deaths_confirmed_clean'] = h1n1_dataframe['Deaths_confirmed_clean'].str.replace(',','').asty
pe(float)
h1n1_dataframe.head()
```

Out[22]:

	Country	Cases_confirmed_clean	Deaths_confirmed_clean	total_cases
0	United States	113690.0	3433.0	1467884.0
1	Brazil	58178.0	2135.0	233142.0
2	India	33783.0	2024.0	90927.0
3	Mexico	70715.0	1316.0	47144.0
4	China	120940.0	800.0	84044.0

Correlation Between H1N1 Deaths and Total Cases

- Plotting Scatter Plot.
- Calculate correlation factor. (Value = 0.799)

Observations:

- There's a high correlation between them and GDP sounds as an effective feature to add for country features.

In [23]:

```
px.scatter(h1n1_dataframe,'Deaths_confirmed_clean','total_cases',title = "COVID-19 2020 Confirmed vs H1N1 2009 Virus Deaths").show()
h1n1_dataframe[h1n1_dataframe.columns[:]].corr()['total_cases'][:]
```

Out[23]:

```
Cases_confirmed_clean     0.185592
Deaths_confirmed_clean    0.799720
total_cases               1.000000
Name: total_cases, dtype: float64
```

Saving H1N1 Features File

In [0]:

```
countries = h1n1_dataframe['Country'].unique()
country_deaths_flu_2009_dict = {}
country_cases_flu_2009_dict = {}
for country in countries:
  country_deaths_flu_2009_dict[country] = h1n1_dataframe[h1n1_dataframe['Country'] == country]['Deaths_confirmed_clean'].to_numpy()[0]
  country_cases_flu_2009_dict[country] = h1n1_dataframe[h1n1_dataframe['Country'] == country]['Cases_confirmed_clean'].to_numpy()[0]
# save file to permenant storage on drive.
save_dict_to_pickle(country_deaths_flu_2009_dict, STORAGE_DIR+"deaths_flu_2009.pickle")
save_dict_to_pickle(country_cases_flu_2009_dict, STORAGE_DIR+"cases_flu_2009.pickle")
```

Summary

We tried other features in this dataset, but with no good correlation numbers thus we will use only those 2 new features to add to our Model Class 2 Dataset.

GDP and H1N1 Virus Cases showed high correlations and increased the Model efficeny by some magnitude.