**1.1 Introduction**

Coronavirus disease (COVID-19) is an infectious disease caused by a virus. Cases of Covid-19 first emerged in late 2019, when a mysterious illness was reported in Wuhan, China. The cause of the disease was soon confirmed as a new kind of coronavirus, and the infection has since spread to many countries around the world and become a pandemic.

This project focus is to analyze the data related to COVID-19 and create models to predict what is going to happen. All the analyses were done in Python and PySpark and codes produced are attached in the end of the document.

Data chosen to perform the analysis was extracted from a github repository operated by the Johns Hopkins University Center (link: <https://github.com/CSSEGISandData/COVID-19>).

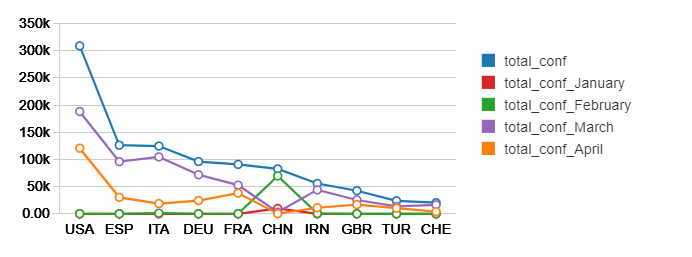
**1.2 Data Overview**

Data extracted contains daily information regarding number of confirmed cases, number of deaths and number of recovered cases summarized by country and in some cases, province/region. First day with available information is January 22nd, 2020.

**1.2.1 Confirmed cases**

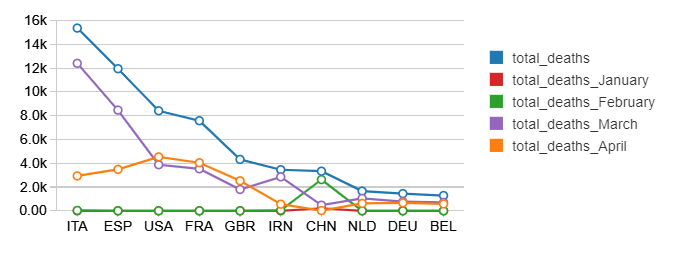
Between January 22nd and April 4th, there were almost 1,2MM cases confirmed around the world. As we can see in the chart below, virus has already spread to most of the countries.

Chart below shows the top 10 countries with most confirmed cases so far (blue line). Looking at the when each country was affected, China, which was the first country to announce cases, had a big number of cases in February, but few cases in March. For the other countries with high volume of cases, such as Spain, no considerable amount of cases was reported in January and February, but a large portion was reported in March and early April.



**1.2.2 Deaths**

In terms of deaths due to the virus, Italy and Spain are the countries that have reported largest number of occurrences so far. Most of them happened in March. However, based on the number of cases observed in the first four days of April, it is expected that total cases in April will surpass last month events.



**1.2.3 Recoveries**

When it comes to recoveries, China has shown majority of the cases already reported. Since it was the first country to report high volume of cases, other countries with recent increase in confirmed cases are expected to show similar recoveries increase in the next coming months.

