UMA MAHESHWARI
umamohantm@gmail.com
+91 9952556206

COVID-19, 2020 ANALYSIS AND PREDICTION OF COVID-19 INFECTED.

July, 2020

Context

Corona virus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome corona virus 2. It was first identified in December 2019 in Wuhan, Hubei China and resulted in an ongoing pandemic. As of 27 July 2020, more than 16.2 million cases have been reported across 188 countries and territories, resulting in more than 647,000 deaths. More than 9.36 million people have recovered. The virus is primarily spread between people during close contact, most often via small droplets produced by coughing, sneezing, and talking. Preventive measures include social distancing, washing hands with soap and water often, sanitizing frequently touched surfaces, wear a face mask, cover coughs and sneezes with a tissue, avoid sharing personal household items, etc,. Social distancing strategies aim to reduce contact of infected persons with large groups by closing schools and workplaces, restricting travel, and cancelling large public gatherings.

This project is divided into two parts:

- This section includes the analysis of history of covid-19 in which countries this pandemic started and how many countries are affected till date, early stage of this pandemic versus today (August 9th, 2020), count of covid-19 infected during the period from December 31st, 2019 to August 9th, 2020, fatality of this pandemic, strictness of lockdown, population density of the countries because it plays a major role in this spread, hand washing facilities and sanitary facilities, hospital facilities , etc,.

- This section deals with the prediction of total covid-19 infected people in the upcoming 10 days by considering the infected graph for the past few months. Here ARIMA model is used for predicting the covid-19 infected people count.

The Data

The data used in this project is a collection of the COVID-19 data maintained by *Our World in Data*. It is updated daily and includes data on confirmed cases, deaths, and testing, as well as other variables of potential interest.

Data Wrangling

Data wrangling is the second step in Data Science methods which makes the data more suitable for further analysis. Data wrangling includes

- Data Collection.
- Data Organization.
- Data Definition.
- Data Cleaning.

Data Collection

This CovidData full dataset is readily available in Github our_world_in_data_covid_19. This data contains details about the covid-19 infected and deaths happened in more than 195 countries and above.

The dataset was in CSV file where the information are collected from International Organization for Standardization, National Government Reports, Oxford COVID-19 Government Response Tracker, United Nations Statistics Division, European Centre for Disease Prevention and Control, United Nations, Department of Economic and Social Affairs, Population Division (2017), World Population Prospects: The 2017 Revision, Global Health Observatory Data, etc.

Data Organization

This step of Data Science involves creating sub folders to ensure the project is in well organized manner. Here folders data, figures and models are created to hold the outputs of further steps of the project.

Data Definition

Data Definition includes defining the data such as column names, data types of the column, description of the column, count of unique values or codes and range of unique values or codes including NAN values

This Dataset includes various information like,

- Continent and country details (ISO-CODE).
- Total COVID-19 cases.
- Total COVID-19 deaths.
- Stringency index (measure of strictness of lockdown).
- Population and Population density.
- Median age.
- GDP per capita
- Poverty Rate.
- Cardiovascular death rate.
- Diabetes prevalence.
- Smoking population.
- Hand washing facilities.
- Hospital facilities.
- Life expectancy
- Total tests
- New Test

Data Cleaning

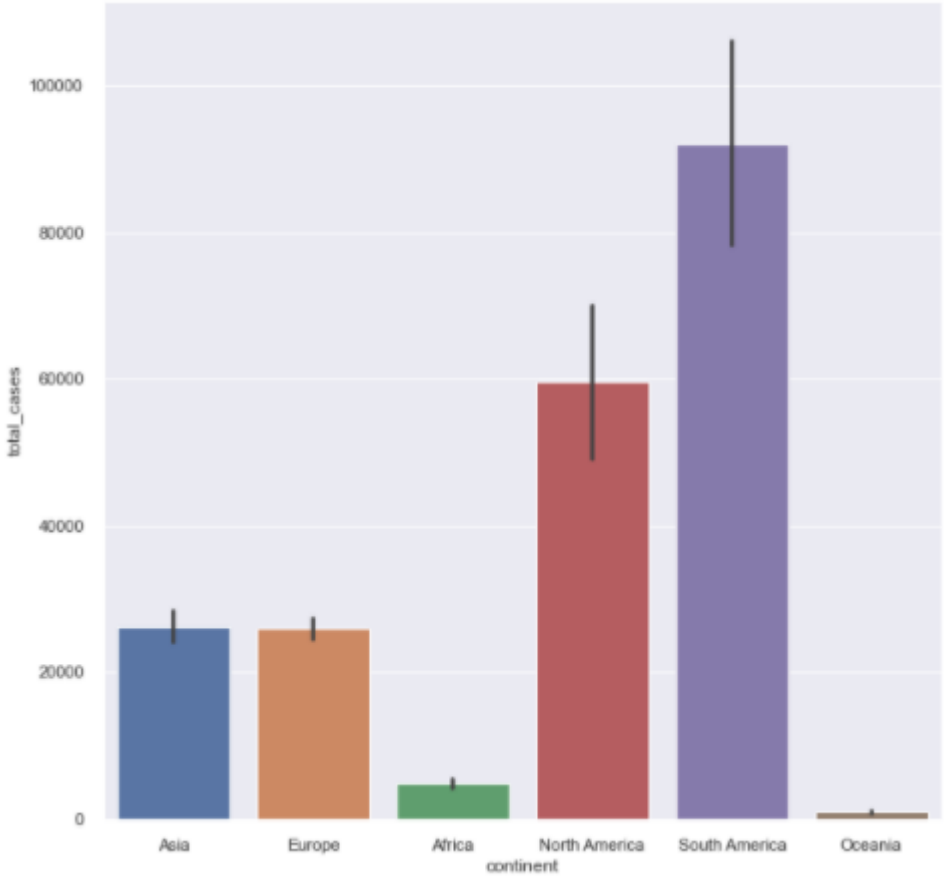
The data originally obtained was in CSV file and directly loaded into the pandas data frame effortlessly. In other words, the dataset we have in our hands is already relatively clean. We will however attempt at learning more about our features and performing appropriate cleaning steps to arrive at a form that is more suitable for analysis. The NAN values are filled with forward fill, backward fill, with mean and median, and interpolate.

Finally all the NAN values are removed and made sure the data is clean and suitable for further process.

Exploratory data visualizations and analysis

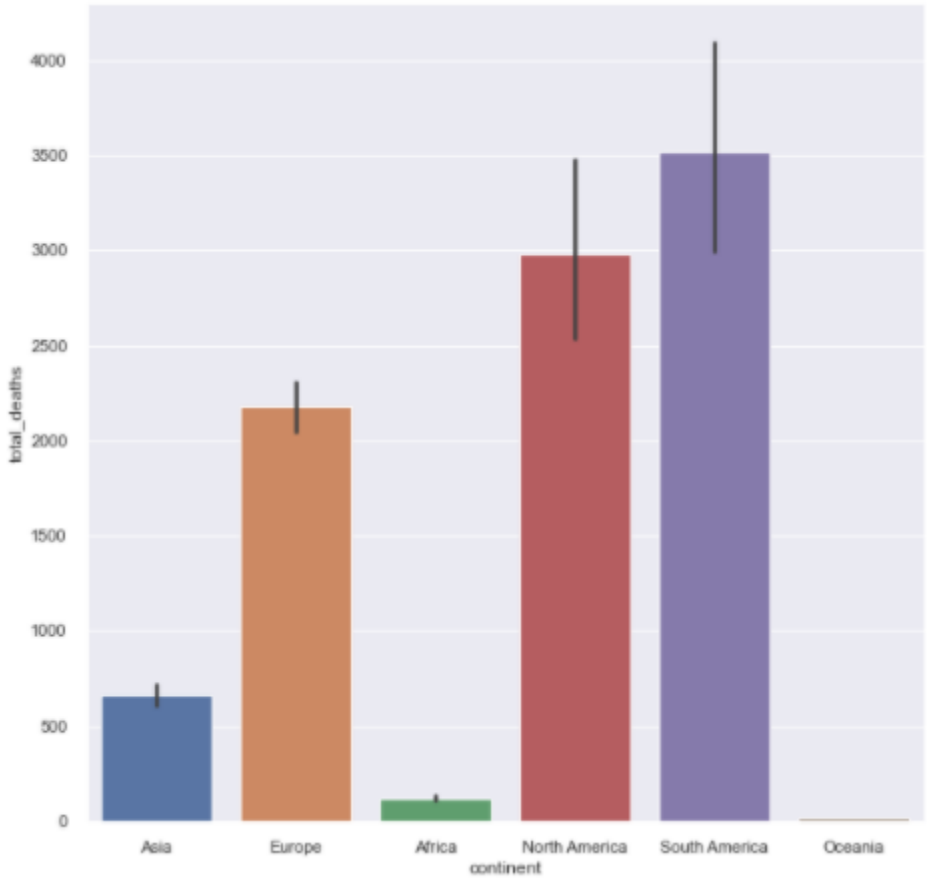
This section involves the visualization part to know the insights from the dataset. This section comprises the first part of this project.

COVID-19 INFECTED CASES ACROSS CONTINENT



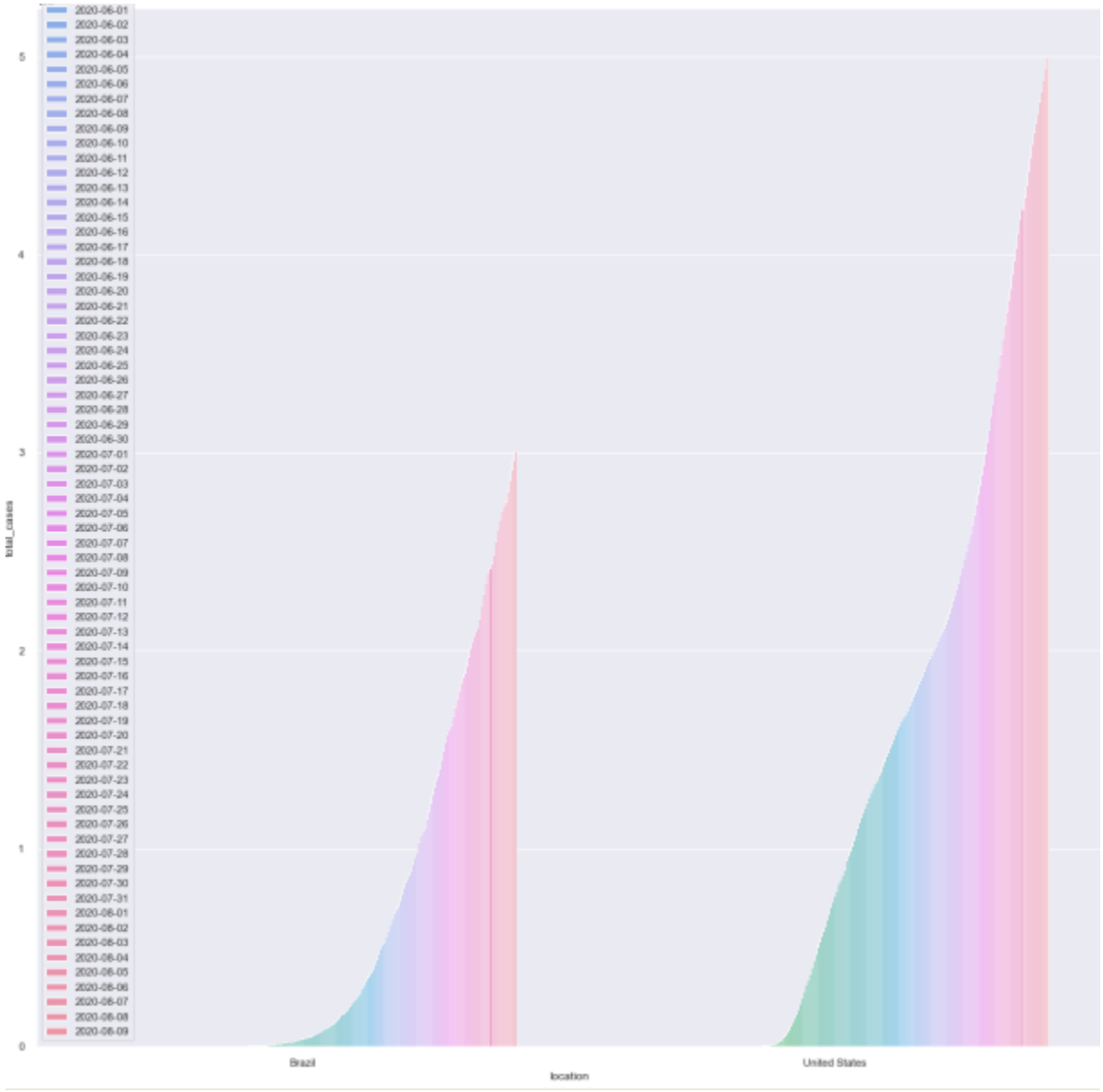
- South America has more number of cases and next to it stands North America.
- Africa as a under developed country and Oceania as a developed country have the least number of cases.
- Both Asia and Europe have approximately same number of people affected with COVID-19.

COVID-19 FATALITIES ACROSS CONTINENT



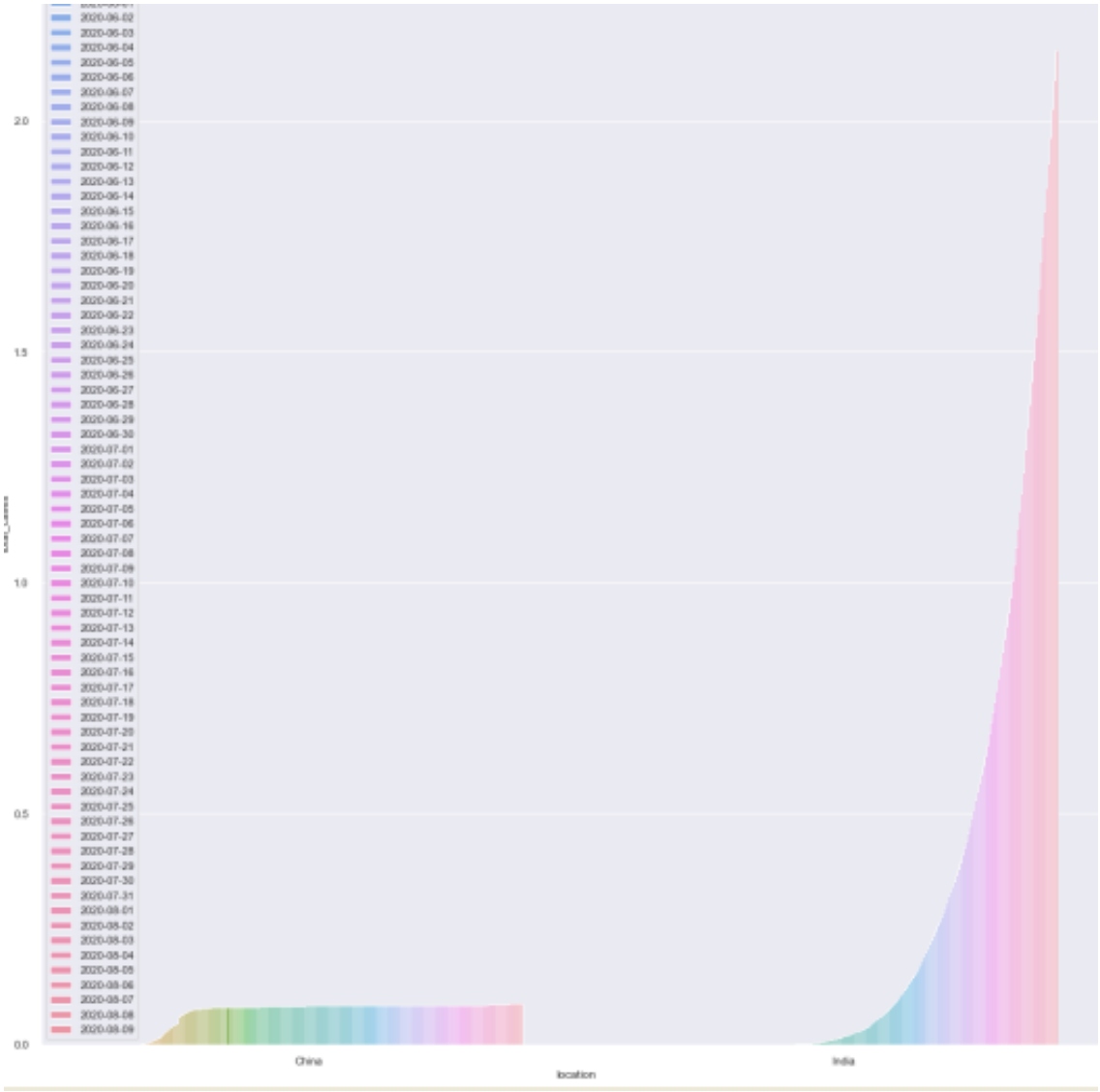
- Same as the total number of cases South America have more number of deaths and next to it stands North America. When we do compare the infected graph with this fatality graph, we could observe the difference in deaths is slightly lesser than the difference in infected count.
- Both Asia and Europe having the same amount of covid-19 cases the death rate is observed unambiguously higher in Europe than in Asia. Both countries having same no of covid-19 infected cases, it's very shocking to see this graph having more deaths in Europe than Asia
- Africa and Oceania have the least death rate. In Oceania almost the death rate is not visible.

US and Brazil COvid-19 Infected Cases



- From the above image we can observe that the first confirmed cases of covid-19 in United States were reported earlier than US and the count drastically increased. And first covid-19 confirmed cases in Brazil were reported later. Comparing to United States the count is less, though it also has drastic increase in covid-19 case count.

China and India COVID-19 Infected Cases



- The outbreak was first identified in Wuhan, China, in December 2019. From the above bar plot we can observe the same. Even though it started very early in China we can observe a flattened curve. The first covid-19 case in India was reported in March, 2020 and had a drastic increase in total numbers of cases when compared with China.