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

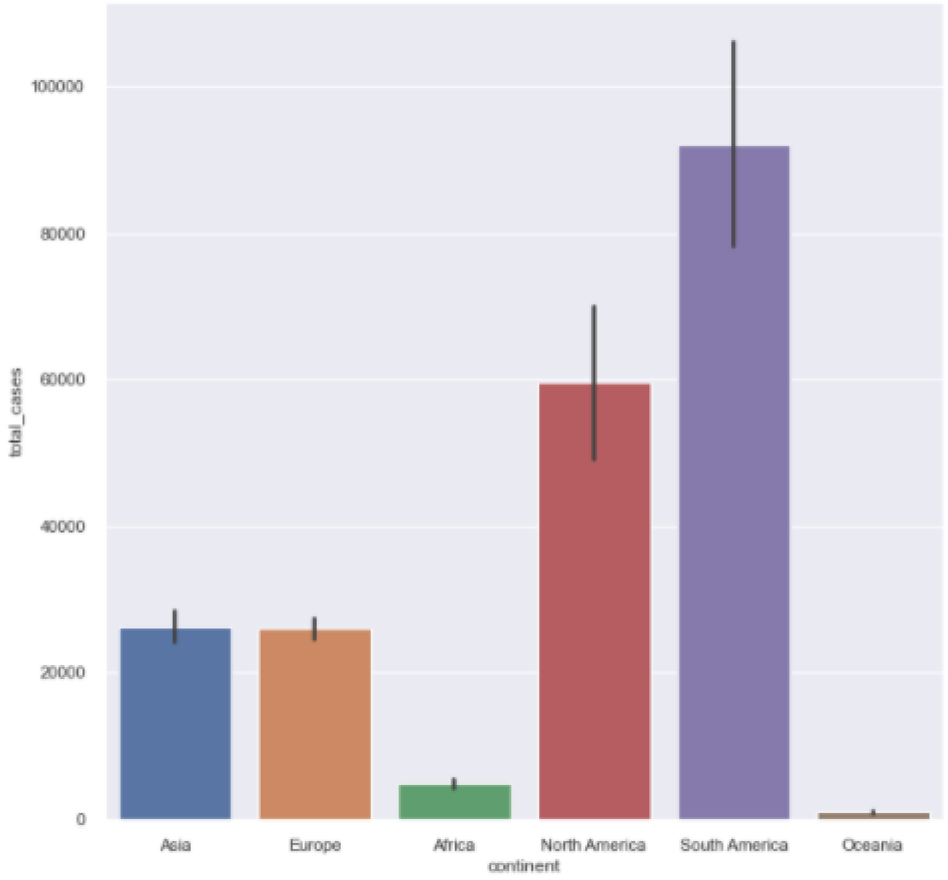
The data originally obtained was in CSV file and directly loaded into the pandas data frame effortlessly. In other words, the dataset we have in our hands is already relatively clean. We will however attempt at learning more about our features and performing appropriate cleaning steps to arrive at a 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

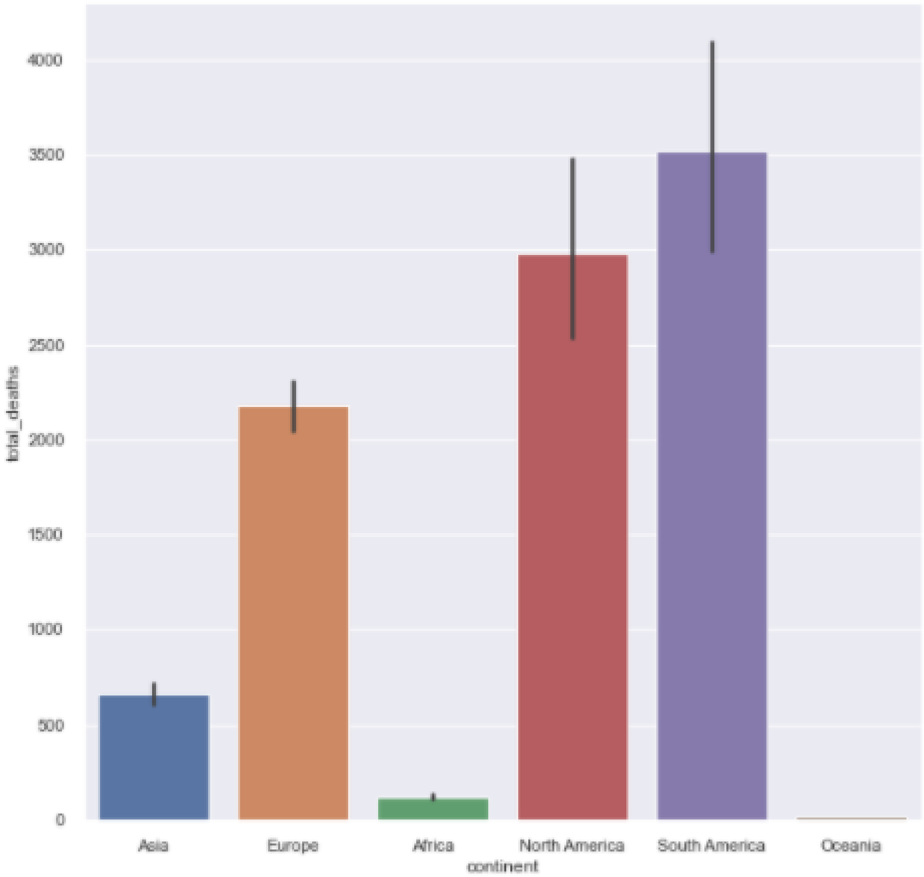
This section involves the visualization part to know the insights from the dataset. This section comprises the first part of this project.

COVID-19 INFECTED CASES ACROSS CONTINENT



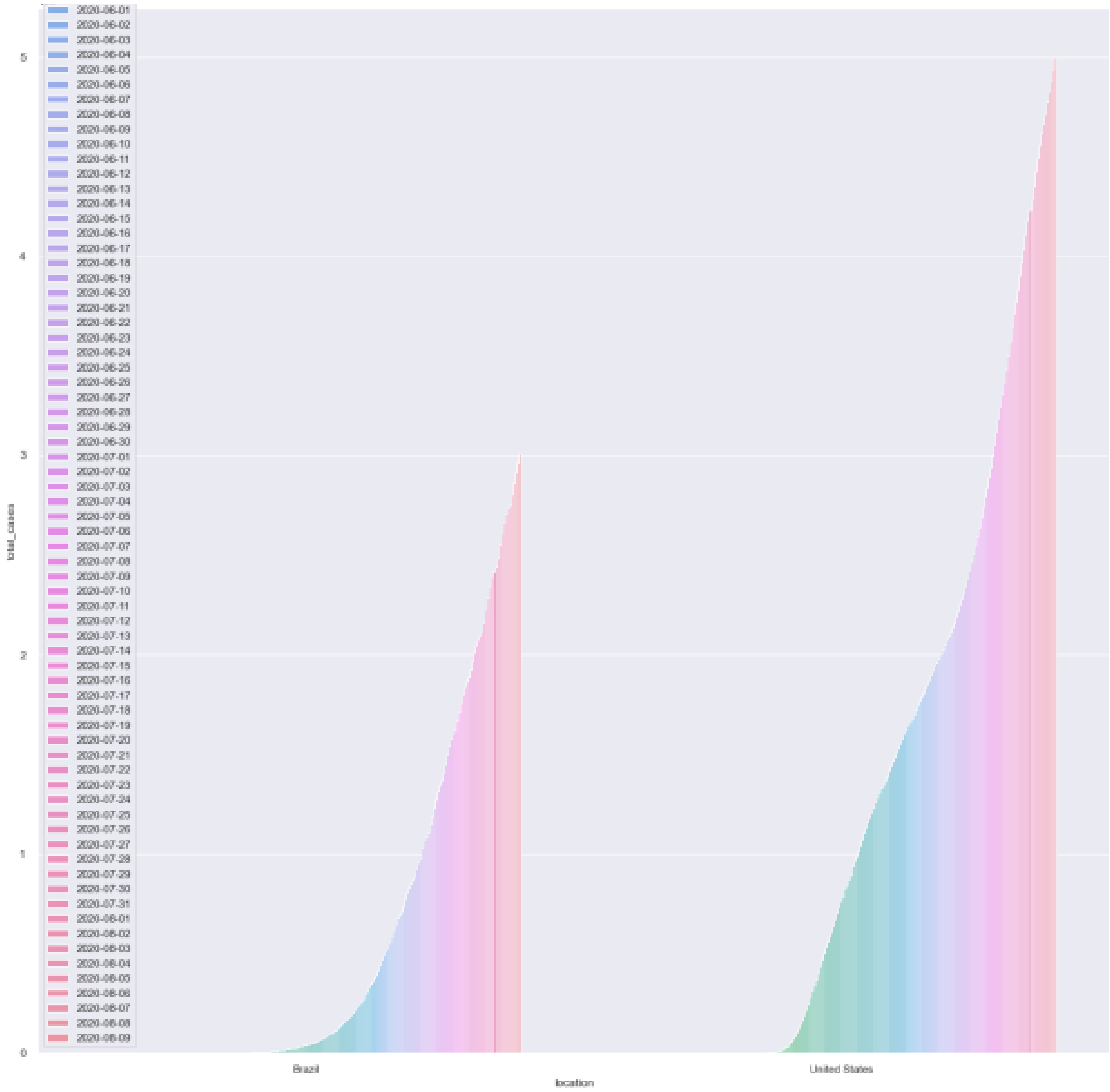
- South America has more number of cases and next to it stands North America.
- Africa as a under developed country and Oceania as a developed country have the least number of cases.
- Both Asia and Europe have approximately same number of people affected with COVID-19.

COVID-19 FATALITIES ACROSS CONTINENT



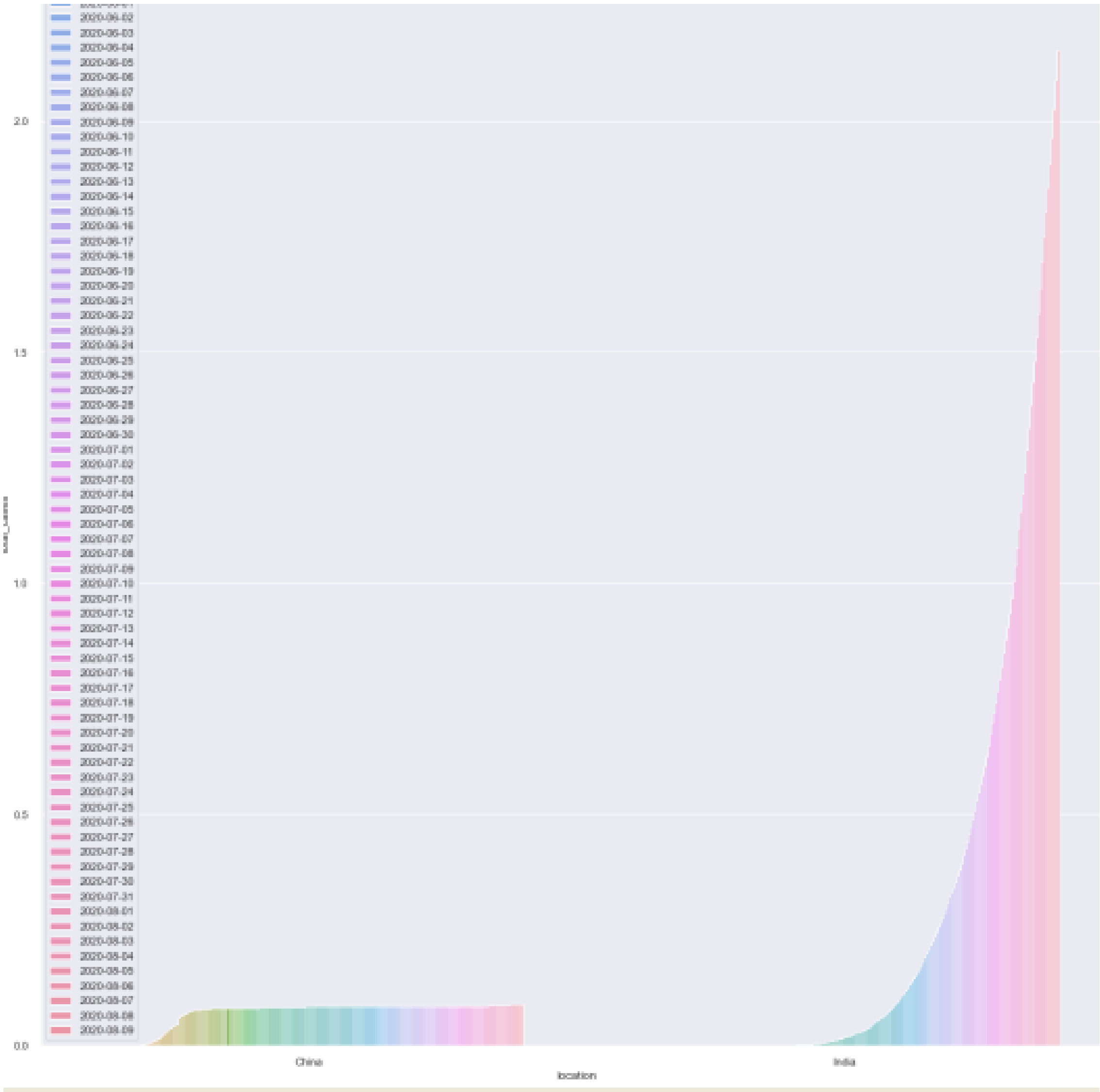
- Same as the total number of cases South America have more number of deaths and next to it stands North America. When we do compare the infected graph with this fatality graph, we could observe the difference in deaths is slightly lesser than the difference in infected count.
- Both Asia and Europe 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



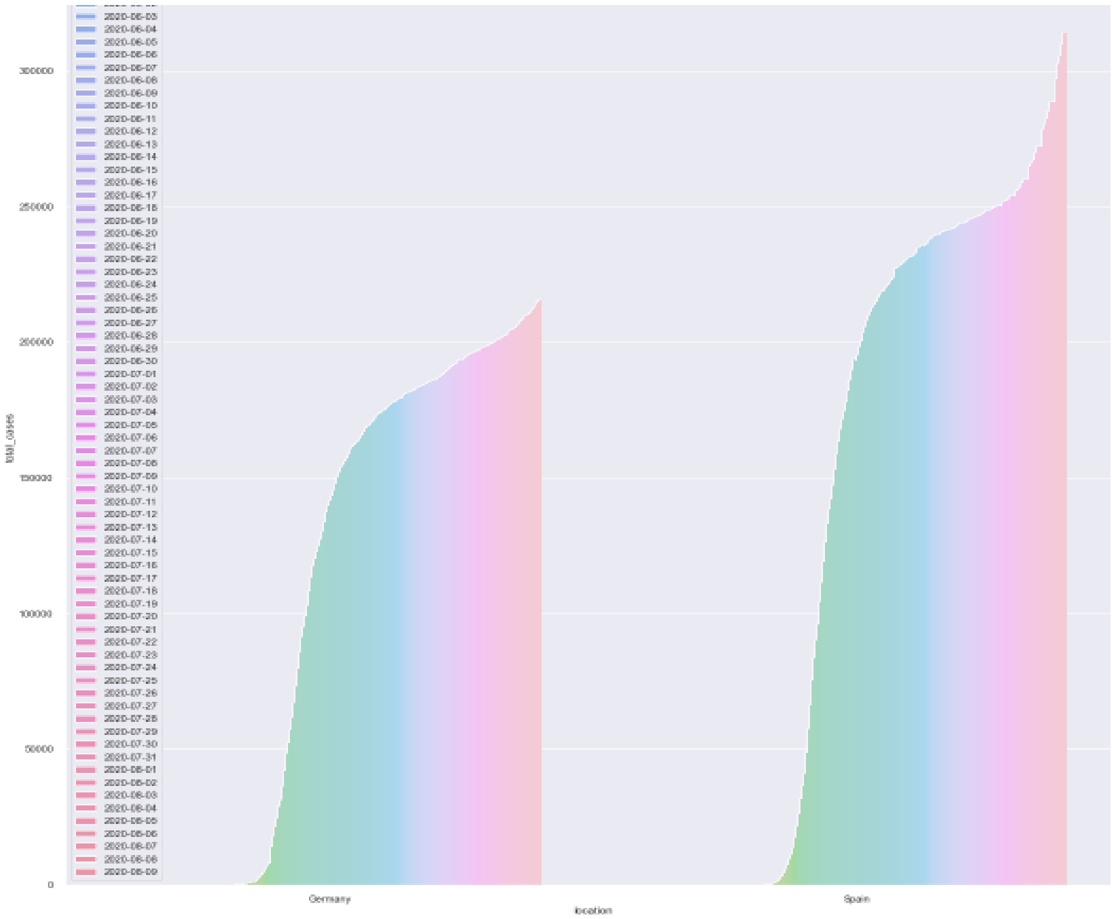
- From the above image we can observe that the first confirmed cases of covid-19 in United States were reported earlier than US and the count drastically increased. And first covid-19 confirmed cases in Brazil were reported later. Comparing to United States the count is less, though it also has drastic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

Germany and Spain COVID-19 Infected Cases



- From the above image we can see that the first covid-19 case was almost reported around the same time both in Germany and Spain.

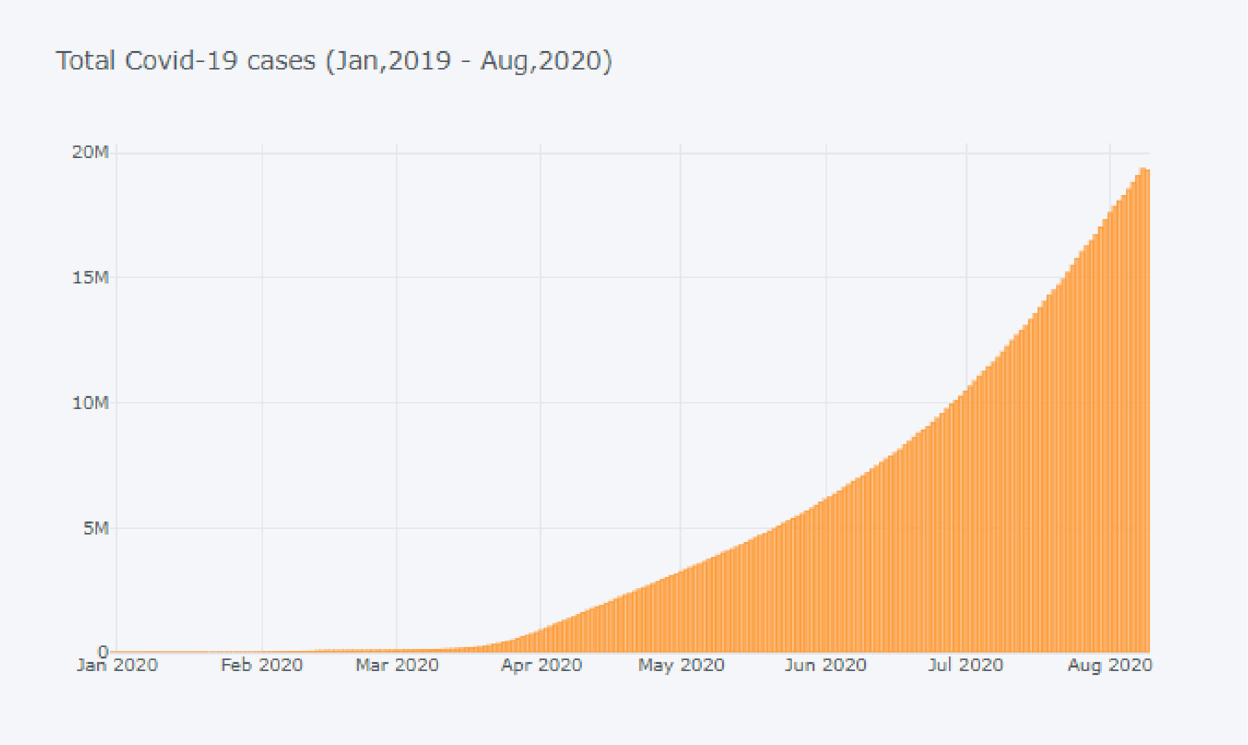
- And we can see a drastic increase in covid-19 infected people counting in both the countries.

- Even though the pandemic started around same time in both countries, Spain has more number of covid-19 infected cases in comparison with Germany.

TOTAL COVID-19 INFECTED CASES

- From this below image we can observe an increasing pattern in total covid-19 infected cases recorded from the month of January to August.

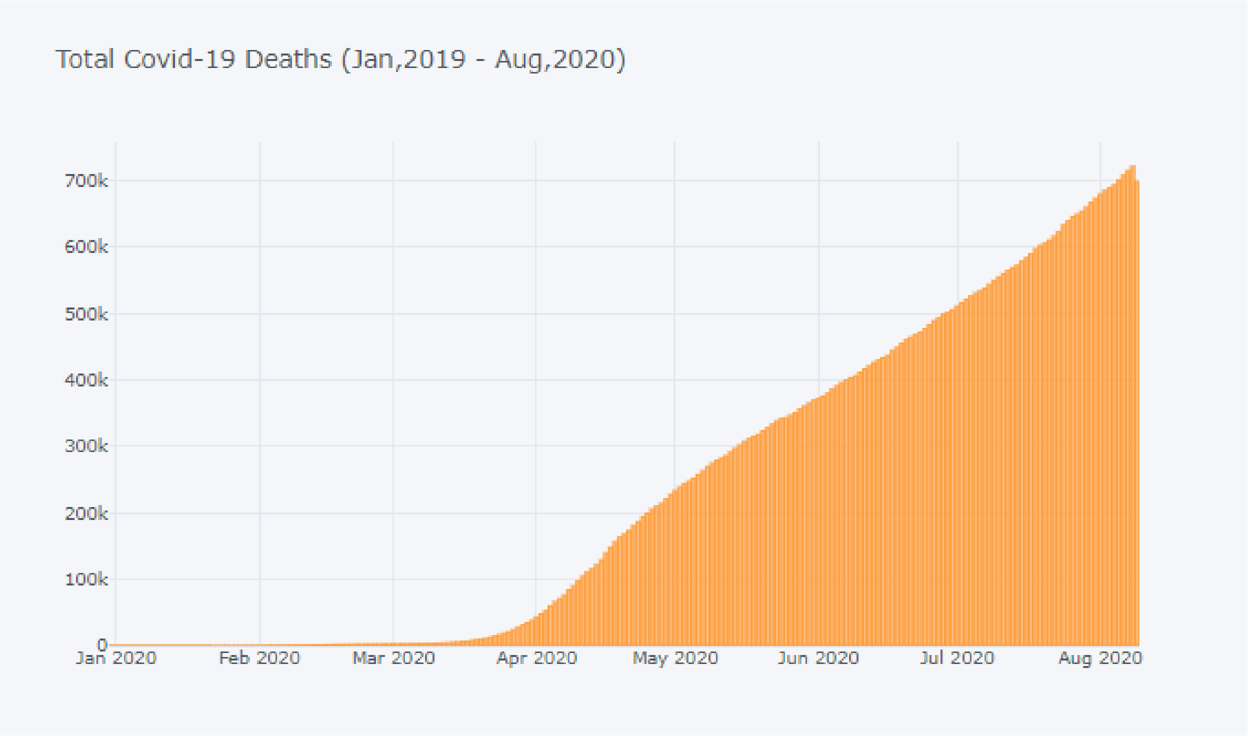
- We can a see a marginal and gradual increasing pattern in the months of January and February, and from the middle of March the cases started to increase drastically.



- On March 1st the total covid-19 cases recorded was 86K and by the middle of the month we can observe the count increased to 160K and by the end of month it almost reached 806K an unanticipated increase.

- And in the next upcoming months the count increased in a drastic manner which is very shocking. And by the end of august the infected count reached almost 18K.

COVID FATALITIES



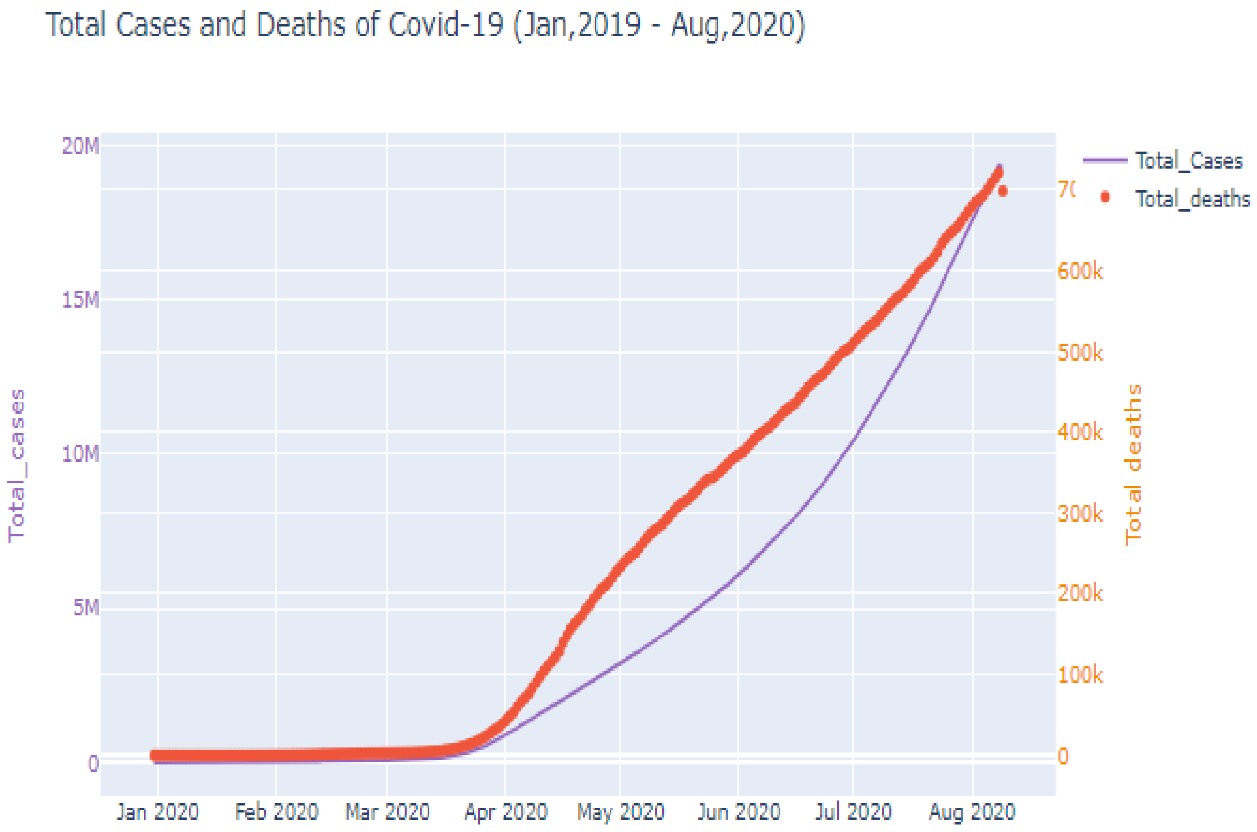
- From the above image we can notice from the middle of March the death started to follow an increasing pattern and by the month of august, as of now the covid-19 death rate almost reached 721K which is very disappointing.

- In the months of January and February the death rate is somewhat less somehow not all the people who are infected are dead even though there is no cure. Somewhat death rate is under control in the first two months of 2020.

- In the start of April there is an increase in the curve somewhat started to lift, and in the middle of April we could see instead of linear curve we almost got a slightly curved line indicating the rapid rise in death count.

- And by the end of August the curve almost reached 700K and this count is very high and alarming.

COVD-19 INFECTED COUNT AND DEATH COUNT

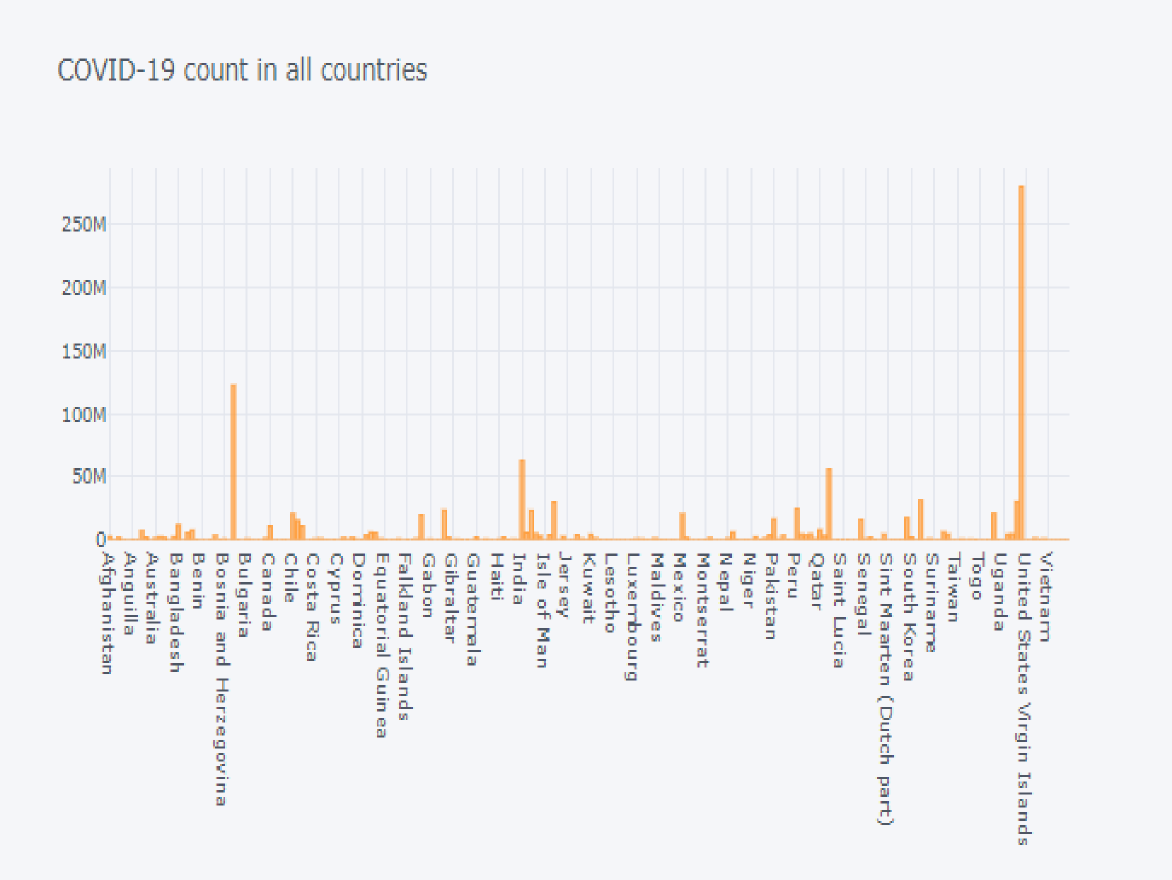


- When we compare both infected and fatality curves, even though the infected is in millions and number of deaths in thousands, we can observe a similar increasing curve pattern in both the lines.

- Almost by mid of March both curves started to have a drastic increase in count. And as a result overall all countries started to have nationwide quarantine and curfews.

- Even though both curve follow the increasing pattern the bend in death rate April mid startled us. The death rate which was following gradual increase started to hike and it's very alarming.

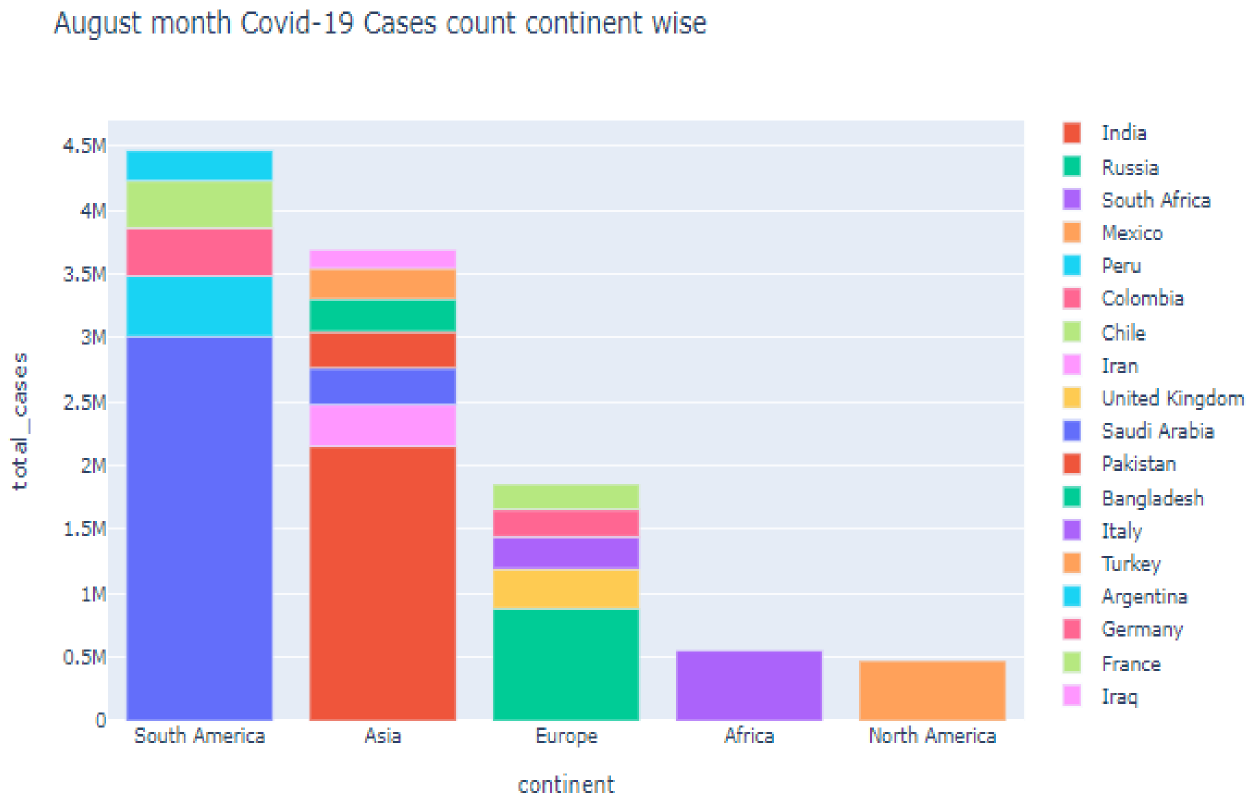
Covid-19 count across countries



- From the above plot we can observe United States has more count almost 280 million covid-19 infected cases and Brazil having almost 122 million covid-19 cases is in second position in covid-19 infected cases as of now.

- India and Russia stands next to that having almost around 65million infected and 60million respectively.

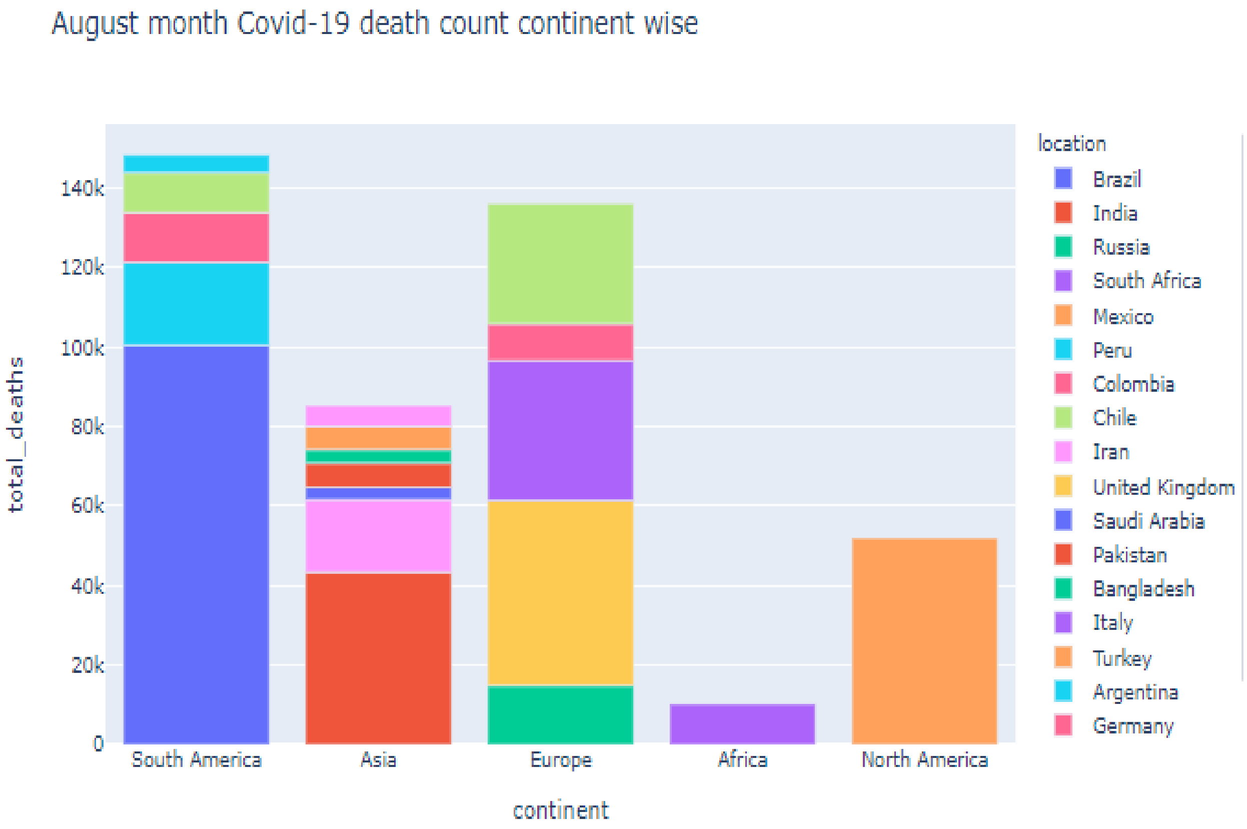
Latest Covid-19 Trend



- In the month of August, South America has the more number of covid-19 cases. South America has almost 235K on August 9th, 2020. Brazil from South America is having almost half of the continent's covid-19 cases (approximately 3 million). The countries which contribute next to Brazil are Peru, Colombia, Chile and Argentina.

- Asia is in second having covid-19 infected people almost more than 3.5 million. Same as Brazil, India holds more than half of the continents' covid-19 infected people (approximately 2.1 million).

- Europe almost has 1.7 million people infected with covid-19 and Russia has more number. Next United States, Italy, Germany and France having 200K to 300K covid-19 cases. Despite of having more cases as a whole, North America tends to have less record of cases in August month. Even though Africa having less total number of covid-19 cases from January till august, Africa have more covid-19 cases recorded than North America lately.

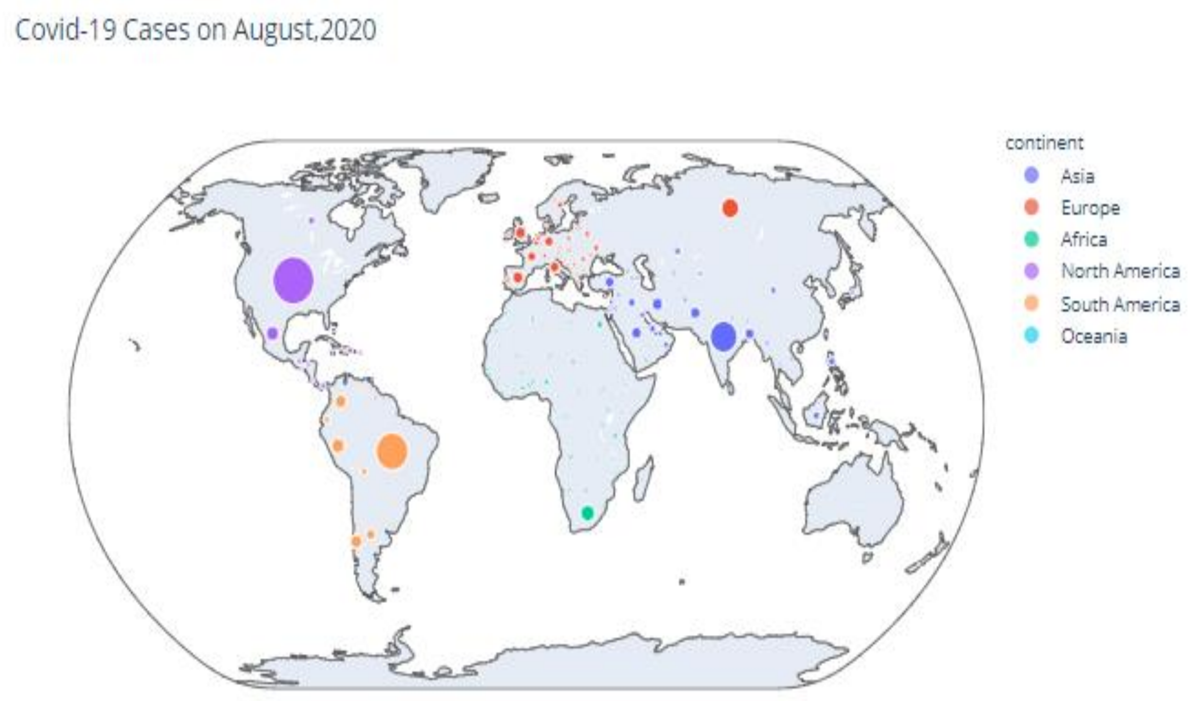


- Again South America is having more covid-19 death rate by August 9th, 2020. It almost reached 140K by August 9th, 2020. Same as covid-19 cases Brazil have more than 100K deaths recorded. And Peru, Chile, Argentina and Columbia have 10K to 20K deaths recorded individually.

- Europe is next to South America having almost 135K covid-19 deaths. Covid-19 cases are more in Russia than in United Kingdom but United Kingdom have more death count than Russia. And Italy and France stands next to United Kingdom. And then Russia and Germany have 14K and 9K death count.

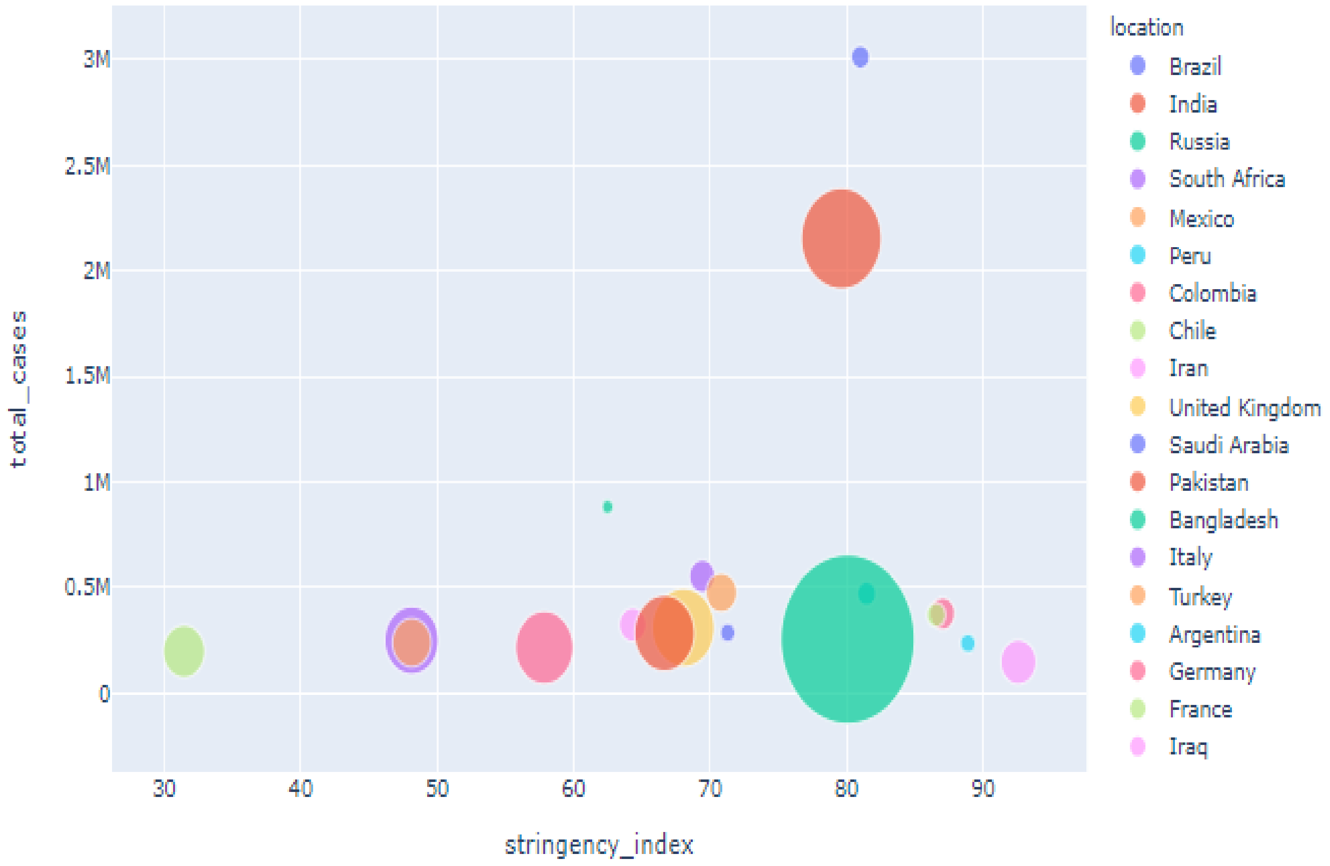
- And Asia has more than 80K deaths in which 50% of death was in India having about 43K deaths. Next Iran has 18k which is approximately 20% from total deaths.

- North America has about 50K and Africa have the least death count of about 10K.



In this image we can see that continent wise covid-19 infected spread.

Stringency Index (measure of Strictness of lockdown)



- Every country's lockdown is different.

- This is a composite measure based on nine response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest). If policies vary at the sub-national level, the index is shown as the response level of the strictest sub-region.

- It is among the metrics being used by the Oxford COVID-19 Government Response Tracker. The Stringency Index is a number from 0 to 100 that reflects these indicators. A higher index score indicates a higher level of stringency.

- The Government Response Stringency Index is a composite measure based on various response indicators including school and workplace closures, stay-at-home policies and travel bans, rescaled to a value from 0 to 100.

- A higher index score indicates a higher level of stringency (100 = strictest response).

- The above plot explains the strictness of lockdown and social distancing in countries having more covid-19 cases lately.

- Brazil having the highest number of covid-19 cases stringency index (a measure of lockdown strictness) has stringency score of about 81.02 which is pretty good but still having lots of infected patients

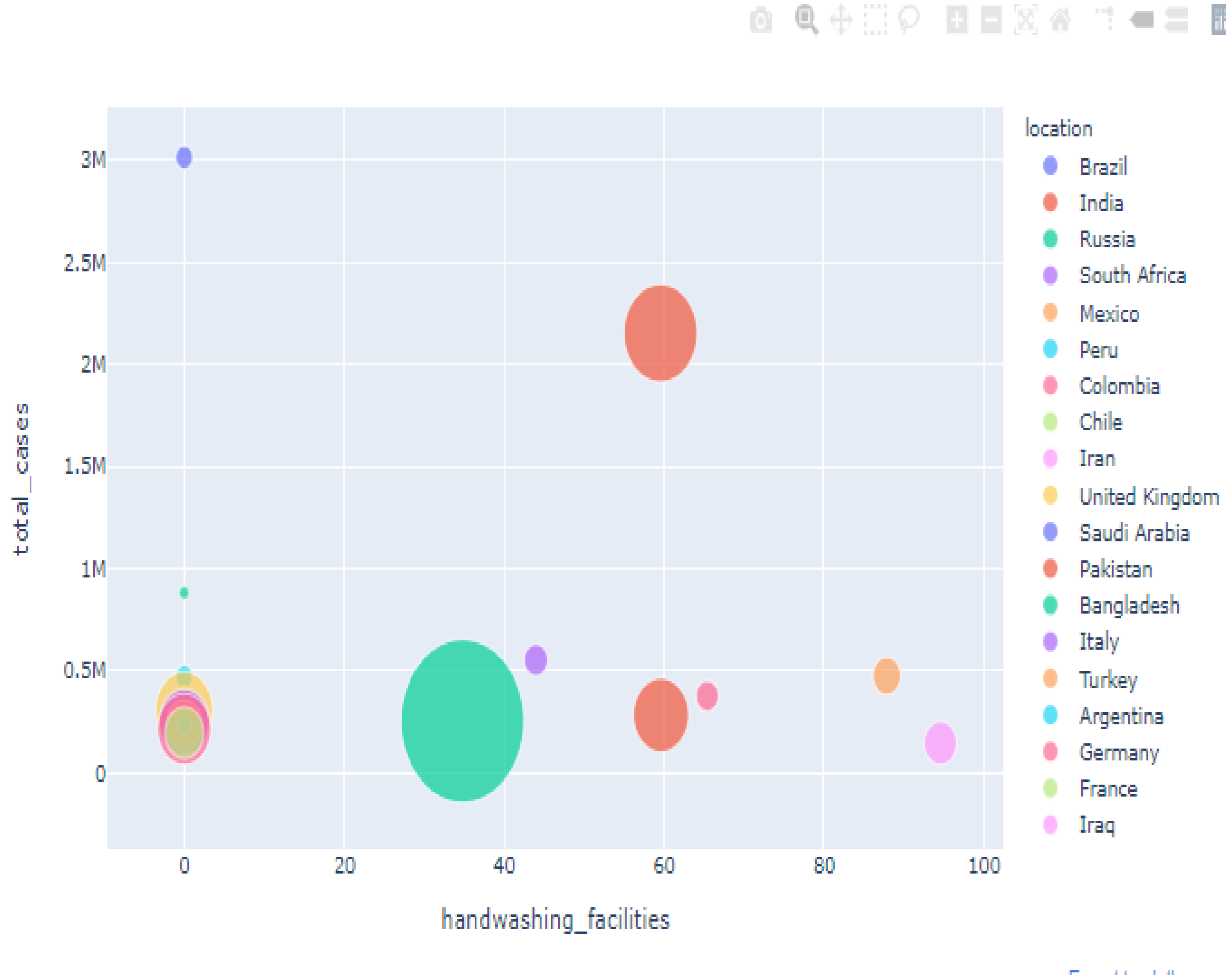
- And next to Brazil stands India having stringency index of 79.63 which is slightly lower than Brazil's.

- And an important fact despite of having more population density than Brazil India have cases less than a million when compared with Brazil.

- Bangladesh with more population density and following the same strictness level like Brazil, the case count is quite good (255K covid-19 infected).

- Russia is in third in having more covid-19 cases and following stringency index of 62 and population density of 8

Hand washing facility Vs COVID-19



- Hands have a crucial role in the transmission of COVID-19. COVID-19 virus primarily spreads through droplet and contact transmission.

- Contact transmission means by touching infected people and/or contaminated objects or surfaces. Thus, your hands can spread virus to other surfaces and/or to your mouth, nose or eyes if you touch them.

- Alcohol-based hand rub products should contain at least 60% alcohol, should be certified and where supplies are limited or cost prohibitive can be made locally by carefully following WHO.

- Plain soap is effective at inactivating enveloped viruses such as the COVID-virus due to the oily surface membrane that is dissolved by soap, killing the virus (Sickbert-Bennett EE et al, “Am J Infect Control” 2005). In addition, hand washing removes germs through mechanical action (WHO Guidelines on Hand Hygiene in Health Care 2009).

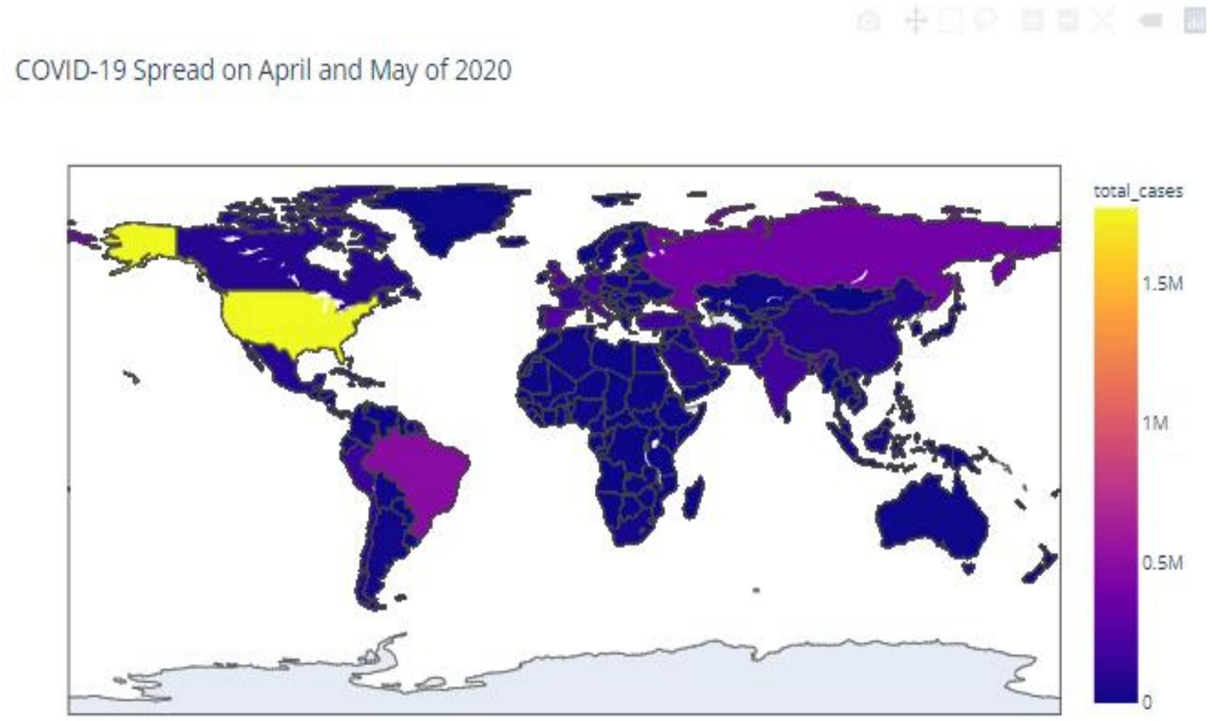
- From this image we could see the hand washing facility is more in Iraq and infected count is less compared to other countries. This graph is mapped with hue =population density.

- And Brazil has low hand washing facility and has more covid-19 infected having very less population density.

- India having hand washing facility of around 60 holds the next position in having more covid-19 infected cases having moderate population density.

- Bangladesh having hand washing facility < 40 have less number of covid-19 infected counts when compared with Brazil and India. And the population density is very high as we could see the bubble size.

- Mexico from South America stands second in hand washing facility (around 85) has minimum number of covid-19 infected.

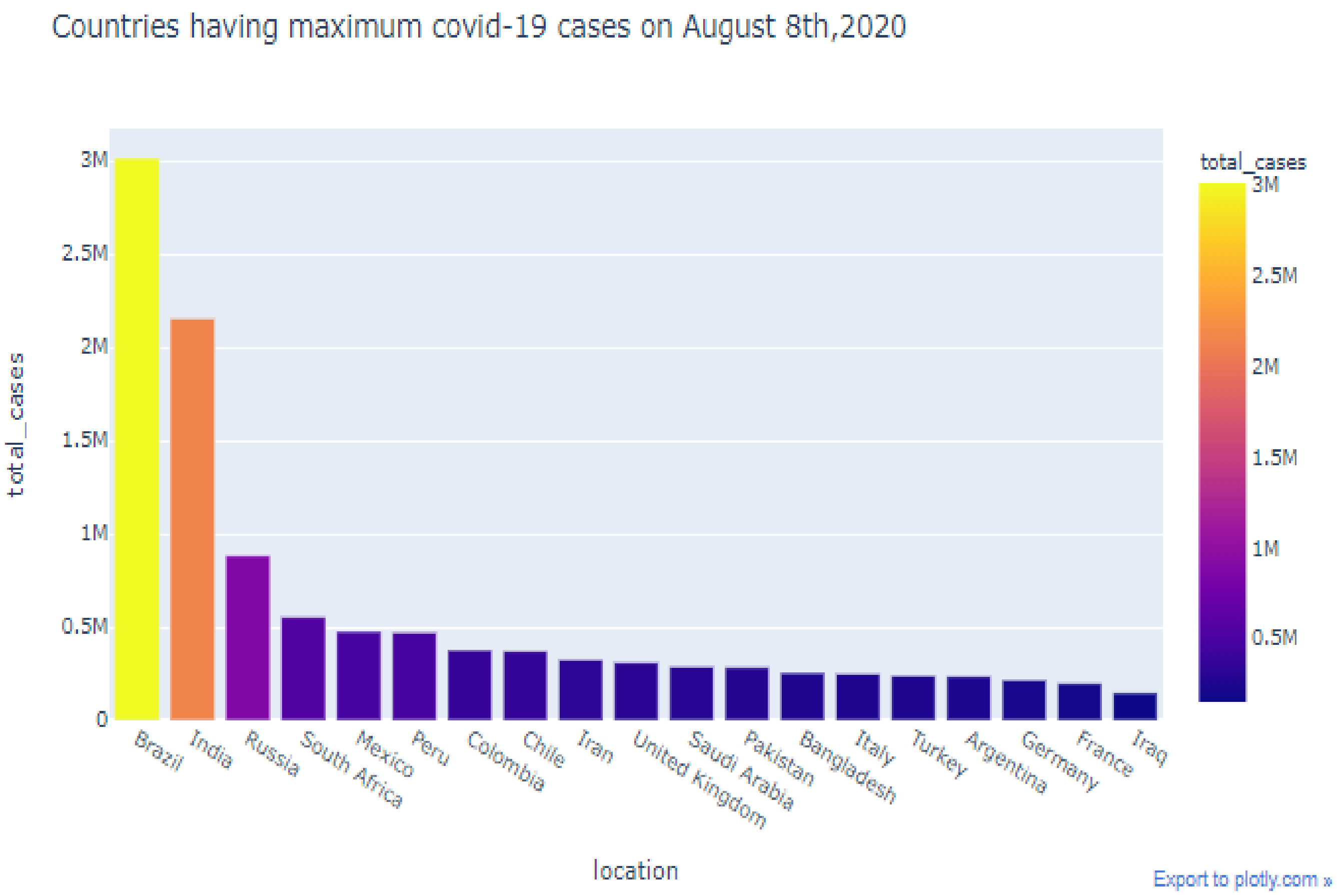


- From the above image we can observe, in the months of April and May US has more number of cases when compared with rest of the world. Clearly we can notice that United States is way ahead of rest of the world. It almost reached more than 1.5 million COVID-19 infected cases.

- India and Iran from Asia, Russia, Germany, Spain, Italy and United Kingdom from Europe and Brazil from South America have COVID-19 infected count in between 0.5million to 1 million.

- Oceania and Africa plays a good role in maintain the COVID-19 infected count. Not even a place from Oceania and Africa crossed 0.5million COVID-19 cases.

Countries having max COVID-19 infected cases (latest)

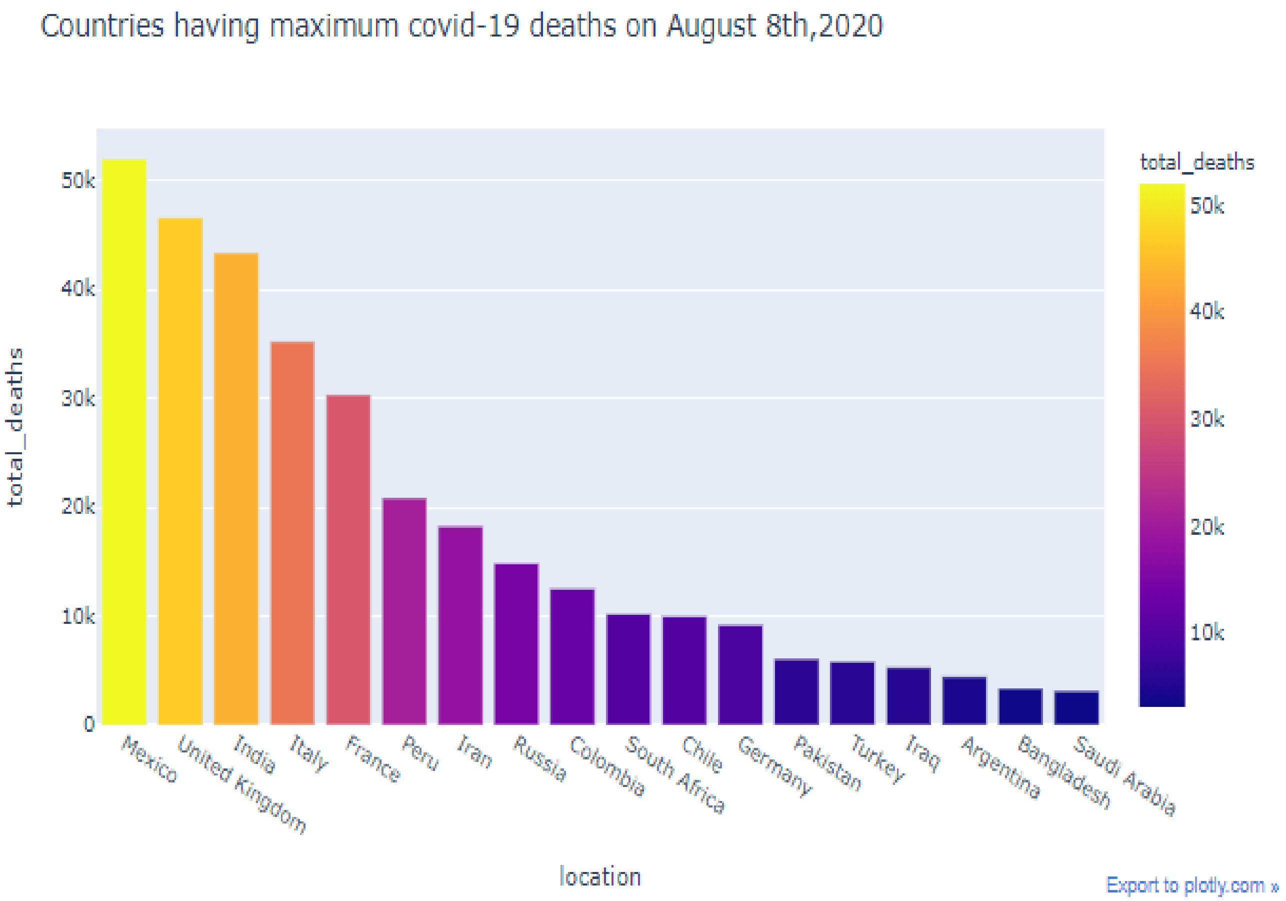


- Brazil and India have the maximum number of covid-19 recorded in the month of August. It’s very startling to see the COVID-19 infected count.

- Brazil has 3 million covid-19 cases and India has 2.1 million COVID-19 infected cases recorded in August month.

- Countries other than these two have less than 1 million covid-19 cases lately. Russia almost has 0.8 million cases and in a way to attain 1 million infected.

Countries having max COVID-19 deaths (latest)



- This image is slightly different than the previous one dealing with total covid-19 infected cases.

- In previous image we observed that Brazil and India shared the highest position of having covid-19 records.

- But Mexico has the higher death rate of about 50K and next to it, United Kingdom holds the second position having 46K deaths recorded in August,2020.

This is the second section in this project: Prediction of COVID-19 infected cases in the upcoming next few days.

Modeling

The approach used in this project is time series analysis and the model applied is ARIMA (Autoregressive Integrated Moving Average). The accuracy is measure using AIC (The Akaike Information Critera).ARIMA has three components – AR (autoregressive term), I (differencing term) and MA (moving average term

AR term refers to the past values used for forecasting the next value. The AR term is defined by the parameter ‘p’ in ARIMA. The value of ‘p’ is determined using the PACF plot.

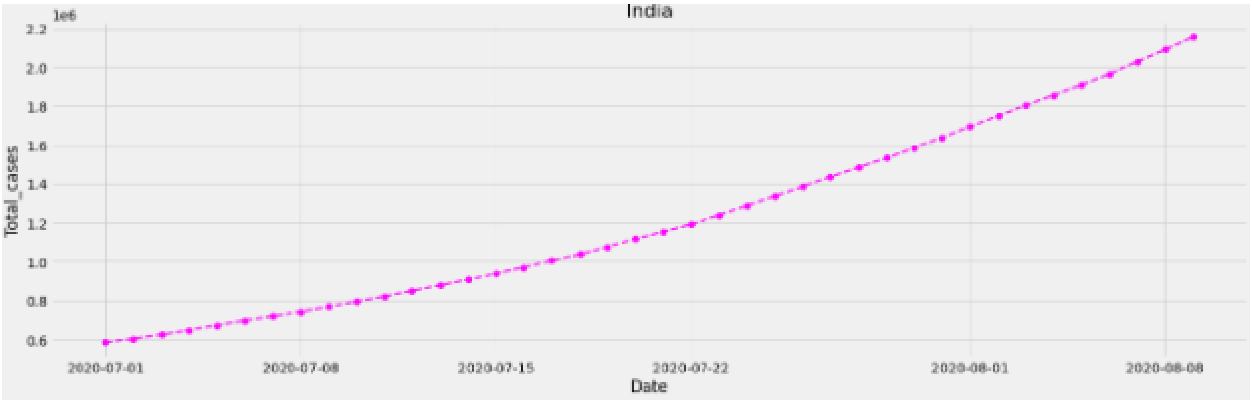
MA term is used to define number of past forecast errors used to predict the future values. The parameter ‘q’ in ARIMA represents the MA term. ACF plot is used to identify the correct ‘q’ value.

Order of differencing specifies the number of times the differencing operation is performed on series to make it stationary.

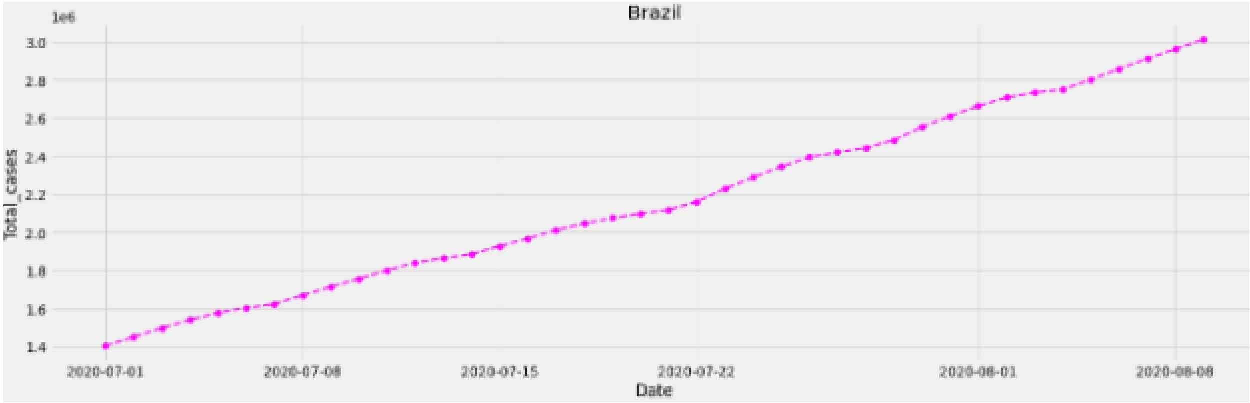
Here using this ARIMA model, the COVID-19 forecasting is done for countries India, Brazil and Russia.

The latest COVID-19 trends in these countries are as follows.

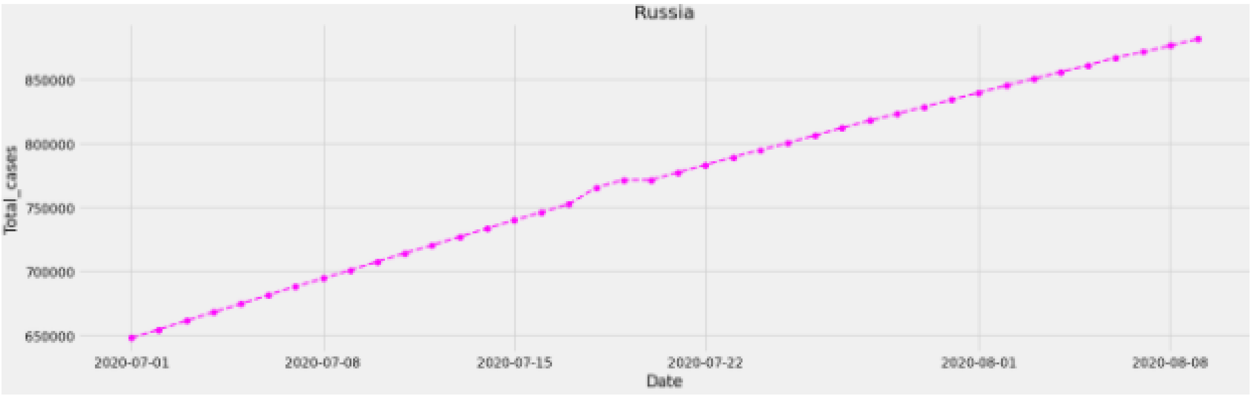
India (from July 1st, 2020 to August 9th, 2020)



Brazil (from July 1st, 2020 to August 9th, 2020)



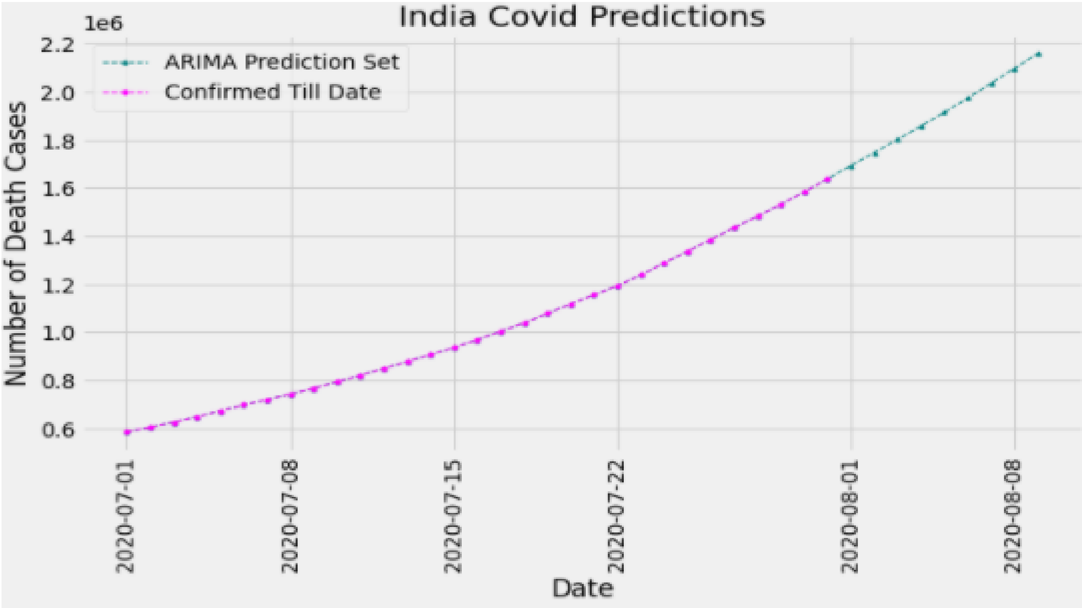
Russia (from July 1st, 2020 to August 9th, 2020)



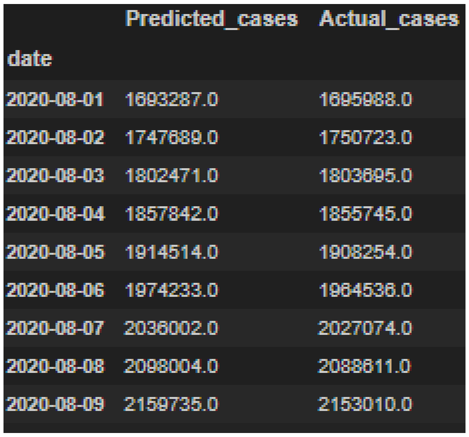
We can observe a linear increasing pattern in three countries COVID-19 infected cases.

The date column from the data frame is converted to date time and made index of the data frame to make it more suitable for modeling.

India COVID-19 prediction

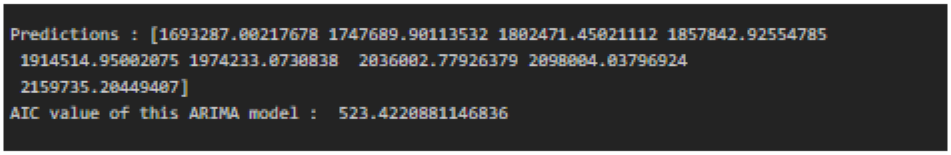


India Predicted Vs Actual



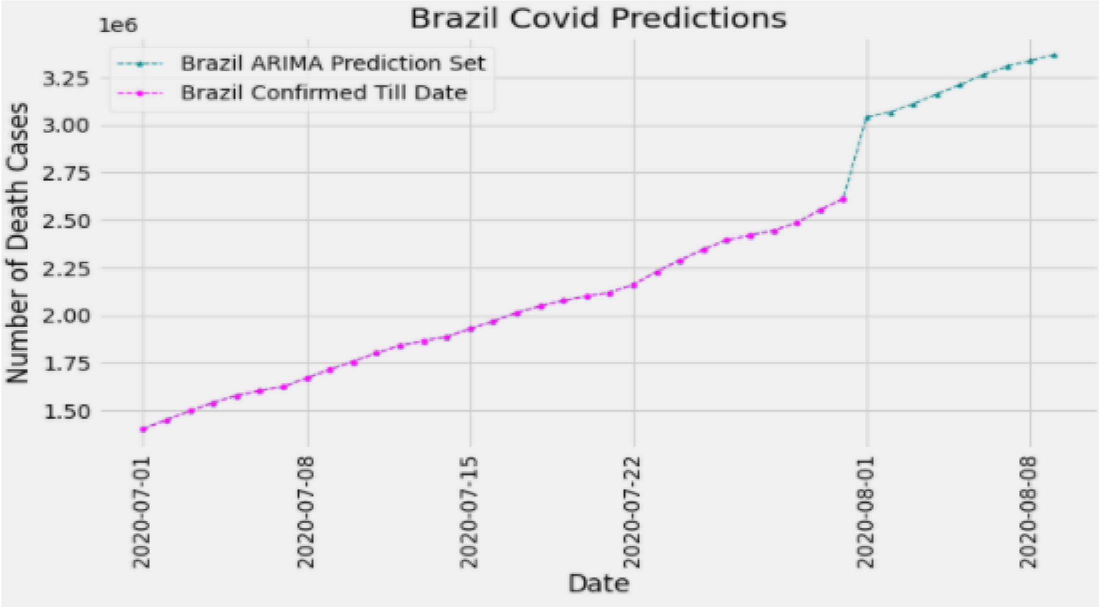
	Predicted_cases	Actual_cases
date		
2020-08-01	1693287.0	1695988.0
2020-08-02	1747689.0	1750723.0
2020-08-03	1802471.0	1803695.0
2020-08-04	1857842.0	1855745.0
2020-08-05	1914514.0	1908254.0
2020-08-06	1974233.0	1964536.0
2020-08-07	2036002.0	2027074.0
2020-08-08	2098004.0	2088611.0
2020-08-09	2159735.0	2153010.0

The metrics AIC (The Akaike Information Critera) is used here to measure the model performance. Lower the AIC value better the model performance. The p, q and d values passed as argument for ARIMA model are 10, 0 and 2 respectively based on partial auto correlation function and auto correlation function graphs.

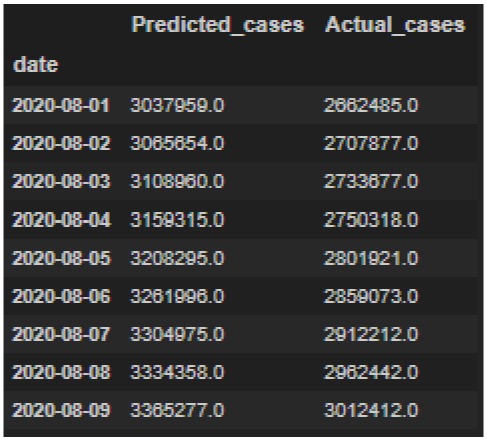


```
Predictions : [1693287.00217678 1747689.90113532 1802471.45021112 1857842.92554785
 1914514.95002075 1974233.0730838  2036002.77926379 2098004.03796924
 2159735.20449407]
AIC value of this ARIMA model :  523.4220881146836
```

Brazil COVID-19 prediction

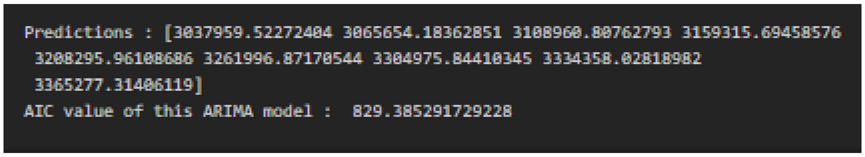


Brazil Predicted Vs Actual



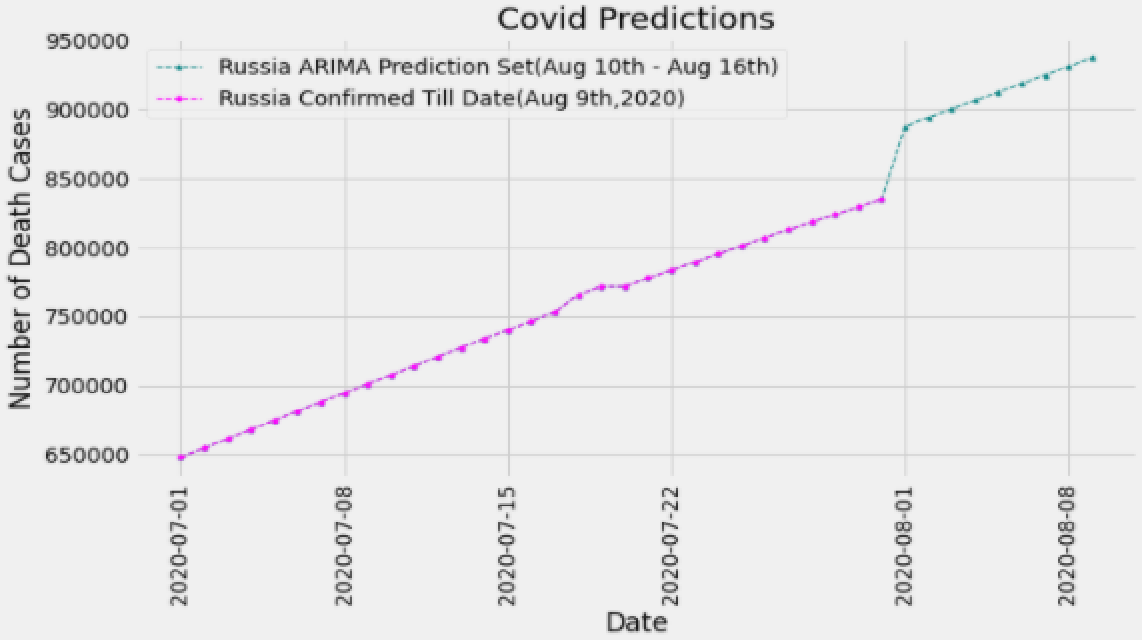
	Predicted_cases	Actual_cases
date		
2020-08-01	3037959.0	2662485.0
2020-08-02	3065654.0	2707877.0
2020-08-03	3108960.0	2733677.0
2020-08-04	3159315.0	2750318.0
2020-08-05	3208295.0	2801921.0
2020-08-06	3261996.0	2859073.0
2020-08-07	3304975.0	2912212.0
2020-08-08	3334358.0	2962442.0
2020-08-09	3365277.0	3012412.0

The p, q and d values passed as argument for ARIMA model are 1, 1 and 1respectively.

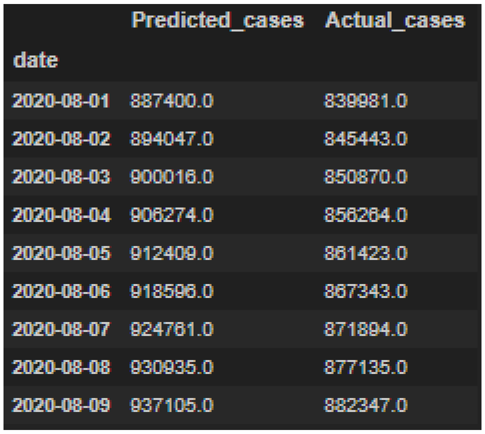


```
Predictions : [3037959.52272404 3065654.18362851 3108960.80762793 3159315.69458576
 3208295.96108686 3261996.87170544 3304975.84410345 3334358.02818982
 3365277.31406119]
AIC value of this ARIMA model :  829.385291729228
```

Russia COVID-19 prediction

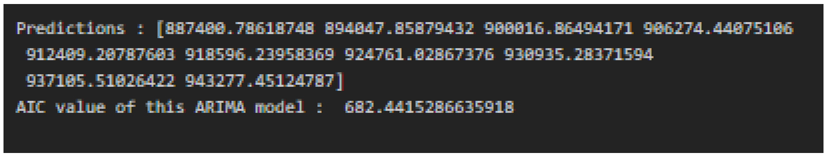


Russia Predicted Vs Actual



	Predicted_cases	Actual_cases
date		
2020-08-01	887400.0	839981.0
2020-08-02	894047.0	845443.0
2020-08-03	900016.0	850870.0
2020-08-04	906274.0	856264.0
2020-08-05	912409.0	861423.0
2020-08-06	918596.0	867343.0
2020-08-07	924761.0	871894.0
2020-08-08	930935.0	877135.0
2020-08-09	937105.0	882347.0

The p, q and d values passed as argument for ARIMA model are 1, 1 and 1respectively.



```
Predictions : [887400.78618748 894047.85879432 900016.86494171 906274.44075106
 912409.20787603 918596.23958369 924761.02867376 930935.28371594
 937105.51026422 943277.45124787]
AIC value of this ARIMA model :  682.4415286635918
```

Conclusion

- Covid-19 has infected over 19 million people worldwide and claimed more than 721K lives with Europe, the United States, Brazil and India overtaking China where the pandemic started last December.

- The drastic outbreak spread is influenced by the countries various factors like strictness of lockdown, population density, hygiene and hand washing facilities, smoking population,

- COVID-19 is still an unclear infectious disease, which means we can only obtain an accurate ARIMA prediction after the outbreak ends.

- This projects helps to be little aware of the kind of situation in the next upcoming days so that we can be prepared to tackle the situation and to help reduce risk.