DATA SCIENCE PROJECT

COVID-19 WORLD ANALYSIS

PROJECT BY:

AKANKSHA YADAV

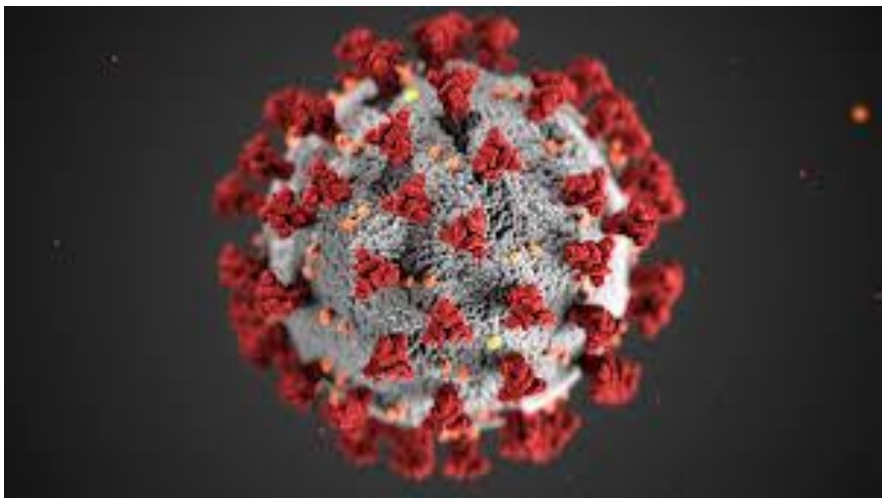
ANUBHAV SHARMA

INDEX

1. INTRODUCTION

Coronaviruses are a group of related RNA viruses that cause diseases in mammals and birds. In humans, these viruses cause respiratory tract infections that can range from mild to lethal. Mild illnesses include some cases of the common cold (which is also caused by other viruses, predominantly rhinoviruses), while more lethal varieties can cause SARS, MERS, and COVID-19. There are as yet no vaccines or antiviral drugs to prevent or treat human coronavirus infections.

Coronavirus is a large family of viruses that can infect animals or humans. In humans, several strains of viruses are known to cause respiratory infections ranging from the common cold to severe diseases such as the Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). The most recently discovered strain is called SARS-CoV-2 strain that is causing COVID19 as it is similar to the SARS-CoV strain that had caused the SARS outbreak.



COVID-19 is a disease, caused by a new strain of corona virus. ‘CO’ stands for corona, ‘VI’ for virus, and ‘D’ for disease. Formerly, this disease was referred to as ‘2019 novel coronavirus’ or ‘2019-nCoV.’ The COVID-19 virus is a new virus linked to the same family of viruses as Severe Acute Respiratory Syndrome (SARS) and some types of common cold.

2. PROJECT AIM

In this project, we intent to analyze about COVID-19 cases all over the world. We would visualize about different countries having active, confirmed, recovered and death figures. We also visualized about death rate for many diseases and also the age group which is most affected by COVID-19.

OUR MAIN AIM: For this project, our main aim is to predict number of new cases all over the world for upcoming 7 days.

3. DATA ANALYSIS

DATA ACQUISITION

It is the process of gathering, filtering and clearing data.
Importing essential libraries is an important step in data acquisition.
Now data was gathered from various sources such as github.

The Pandas library is a useful tool that enables us to read various datasets into dataframe. Using this library, the csv files were converted into data frames.

The main dataset we chose, shows the all the active, recovered, deaths and confirmed cases of the world along with their latitude and longitude. There are many irrelevant columns in the dataset. So removing the irrelevant columns as it will be of no use is necessary. This is known as filtering of data.

	Country_Region	Lat	Long_	Confirmed	Deaths	Recovered	Active
0	Australia	-25.000000	133.000000	11235.0	116.0	8117.0	3002.0
1	Austria	47.516200	14.550100	19439.0	711.0	17335.0	1393.0
2	Canada	60.001000	-95.001000	111317.0	8878.0	98155.0	4285.0
3	China	30.592800	114.305500	85314.0	4644.0	80018.0	652.0
4	Denmark	56.263900	9.501800	13374.0	611.0	12410.0	353.0
5	Finland	61.924100	25.748200	7301.0	328.0	6880.0	93.0
6	France	46.227600	2.213700	211102.0	30141.0	79161.0	101800.0
7	Germany	51.165691	10.451526	201945.0	9089.0	186900.0	5956.0

Our next datasets show about the confirmed, recovered and death cases day wise, which will be useful for the aim of this project.

	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20
0	Afghanistan	33.93911	67.709953	0	0	0	0	0
1	Albania	41.15330	20.168300	0	0	0	0	0
2	Algeria	28.03390	1.659600	0	0	0	0	0
3	Andorra	42.50630	1.521800	0	0	0	0	0
4	Angola	-11.20270	17.873900	0	0	0	0	0

Above dataframe shows confirmed cases day wise of all the countries

	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	1/28/20
0	Afghanistan	33.93911	67.709953	0	0	0	0	0	0	0
1	Albania	41.15330	20.168300	0	0	0	0	0	0	0
2	Algeria	28.03390	1.659600	0	0	0	0	0	0	0
3	Andorra	42.50630	1.521800	0	0	0	0	0	0	0
4	Angola	-11.20270	17.873900	0	0	0	0	0	0	0

5 rows × 180 columns

Above dataframe shows recovered cases day wise of all the countries

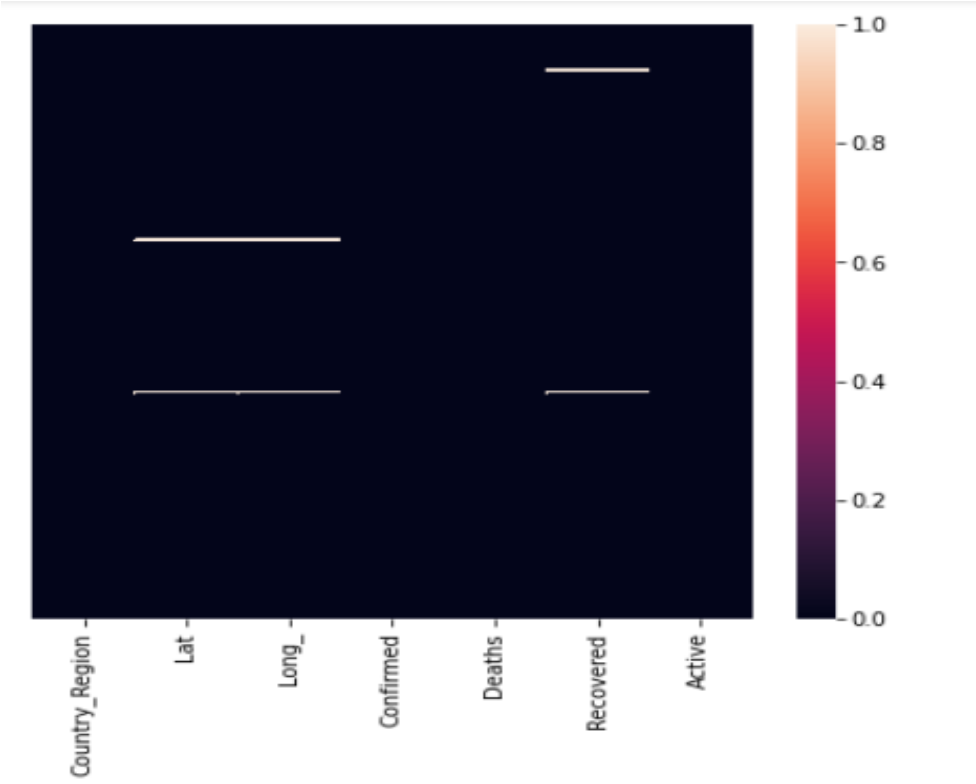
	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	1/28/20	1/29/20	1/30/20
0	Afghanistan	33.93911	67.709953	0	0	0	0	0	0	0	0	0
1	Albania	41.15330	20.168300	0	0	0	0	0	0	0	0	0
2	Algeria	28.03390	1.659600	0	0	0	0	0	0	0	0	0
3	Andorra	42.50630	1.521800	0	0	0	0	0	0	0	0	0
4	Angola	-11.20270	17.873900	0	0	0	0	0	0	0	0	0

5 rows × 180 columns

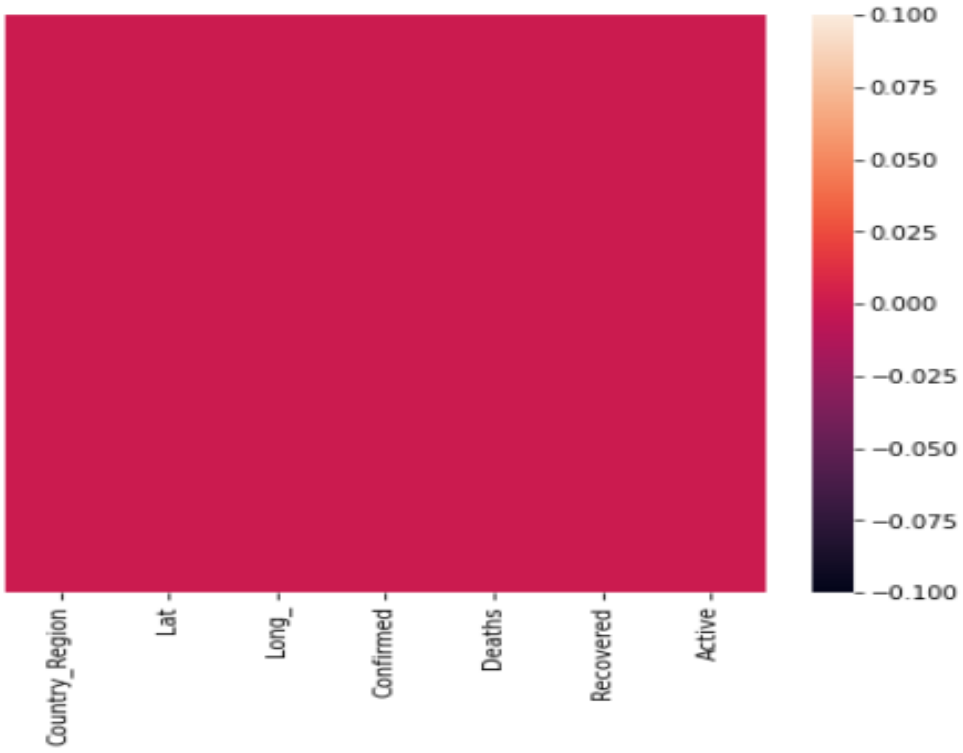
Above dataframe shows number of deaths day wise in all the countries

DATA CLEANING: To fit the model, one needs to get rid of the null values. Hence, the firstly, the columns with null, none or NaN values, were identified. The cleaning of data was done by removing the columns with NaN or null values.

The columns were dropped keeping in mind whether they were really useful for analysis purpose or not. Heat map is generated for visualizing all the null values of the columns for the dataset containing active, recovered, deaths and confirmed cases of all over the world.



Now regenerating heat map again with no null values for verification.



4. EXPLORATORY DATA ANALYSIS(EDA)

It is an approach to analyze data in order to summarize main characteristics of data, gain better understanding of data sets and uncover relationships between different variables.

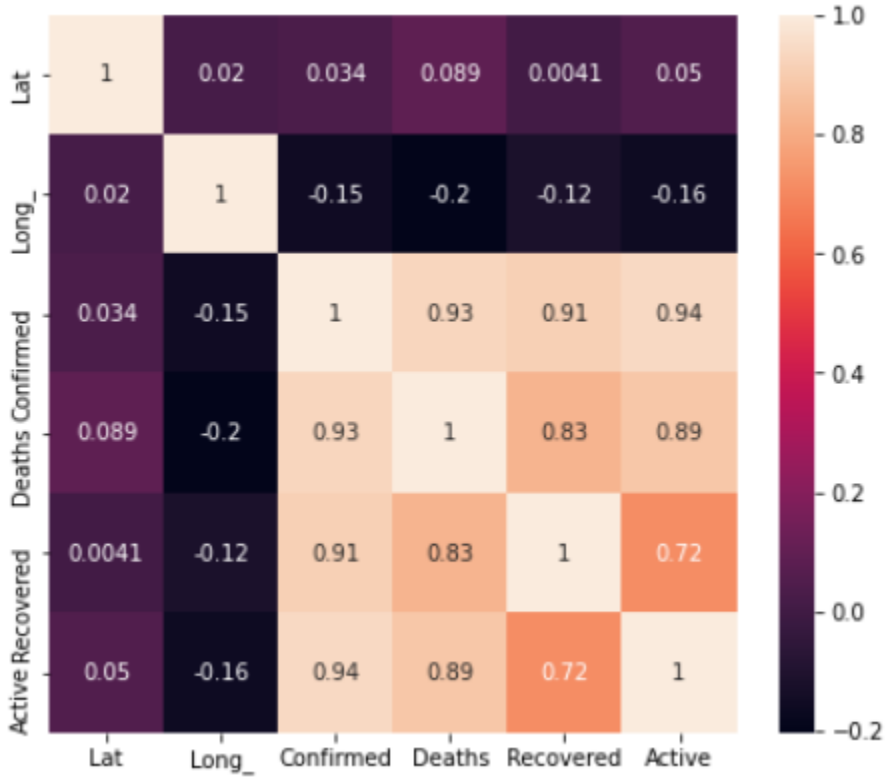
BASIC INSIGHTS OF DATA- To get the concise summary of dataframe, info method is used. It is also used to check whether null values are still present in your dataset or not.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 185 entries, 0 to 187
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Country_Region  185 non-null    object
 1   Lat             185 non-null    float64
 2   Long_           185 non-null    float64
 3   Confirmed       185 non-null    float64
 4   Deaths          185 non-null    float64
 5   Recovered       185 non-null    float64
 6   Active          185 non-null    float64
dtypes: float64(6), object(1)
memory usage: 11.6+ KB
```

DESCRIBE () - It provides us with summary statistics, excluding nan values. It shows the statistical summary of each column, such as count, mean value, standard deviation, etc.

	Country_Region	Lat	Long_	Confirmed	Deaths	Recovered	Active
count	185	185.000000	185.000000	1.850000e+02	185.000000	1.850000e+02	1.850000e+02
unique	185	NaN	NaN	NaN	NaN	NaN	NaN
top	Yemen	NaN	NaN	NaN	NaN	NaN	NaN
freq	1	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	19.980125	17.557432	7.465336e+04	3173.432432	4.204878e+04	2.966248e+04
std	NaN	23.653916	57.535659	3.196369e+05	13079.321687	1.491208e+05	1.851310e+05
min	NaN	-40.900600	-102.552800	1.000000e+01	0.000000	8.000000e+00	0.000000e+00
25%	NaN	5.152149	-7.692100	1.026000e+03	19.000000	5.820000e+02	1.080000e+02
50%	NaN	18.735700	19.699000	5.003000e+03	87.000000	2.430000e+03	1.277000e+03
75%	NaN	40.463667	45.038200	3.485400e+04	668.000000	1.894200e+04	7.637000e+03
max	NaN	64.963100	178.065000	3.606927e+06	138784.000000	1.397531e+06	2.420293e+06

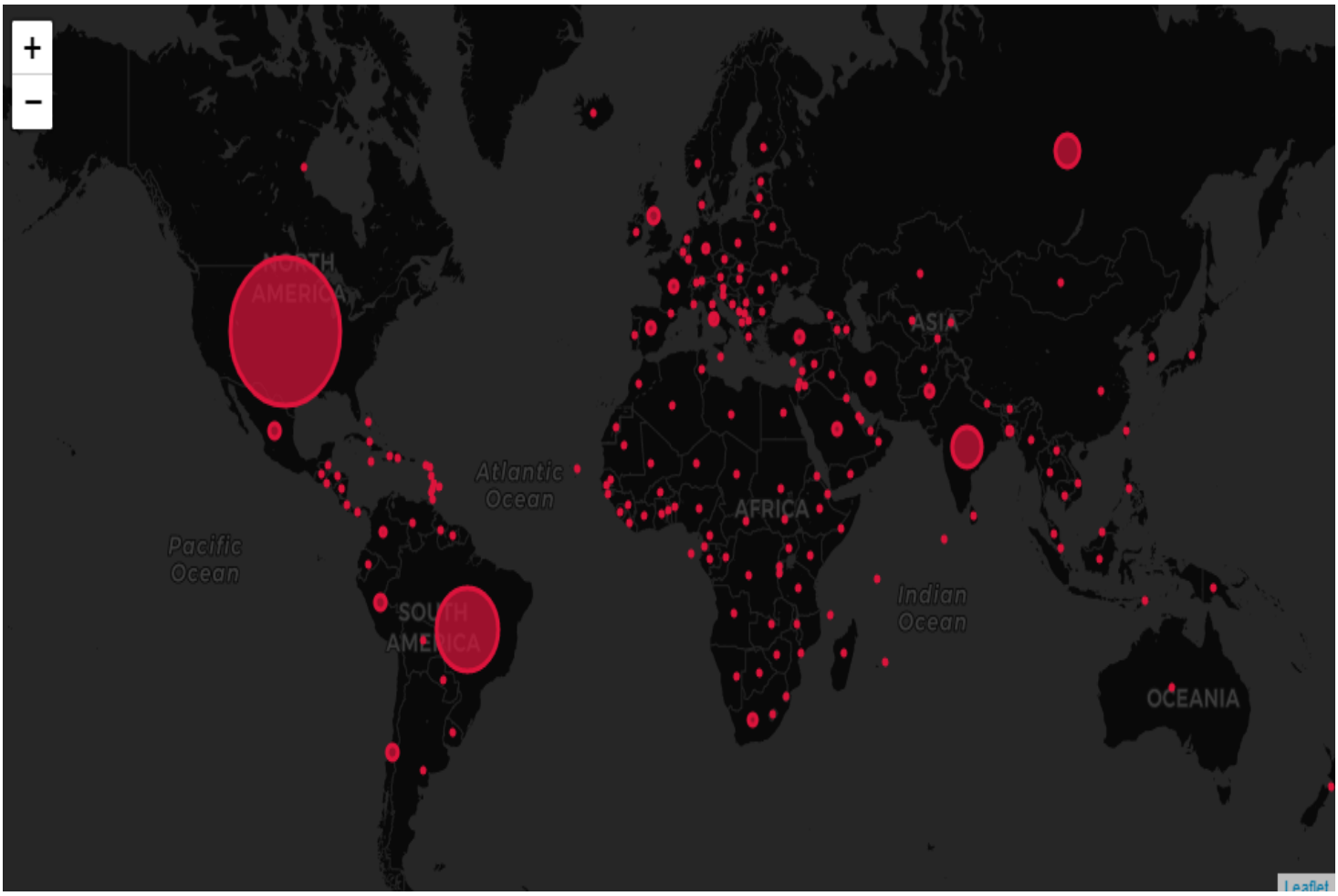
CORRELATION gives us the relationship between each and every variable. Heatmap is a visualization technique to observe correlation between variables in an interesting form.



5.DATA VISUALIZATION

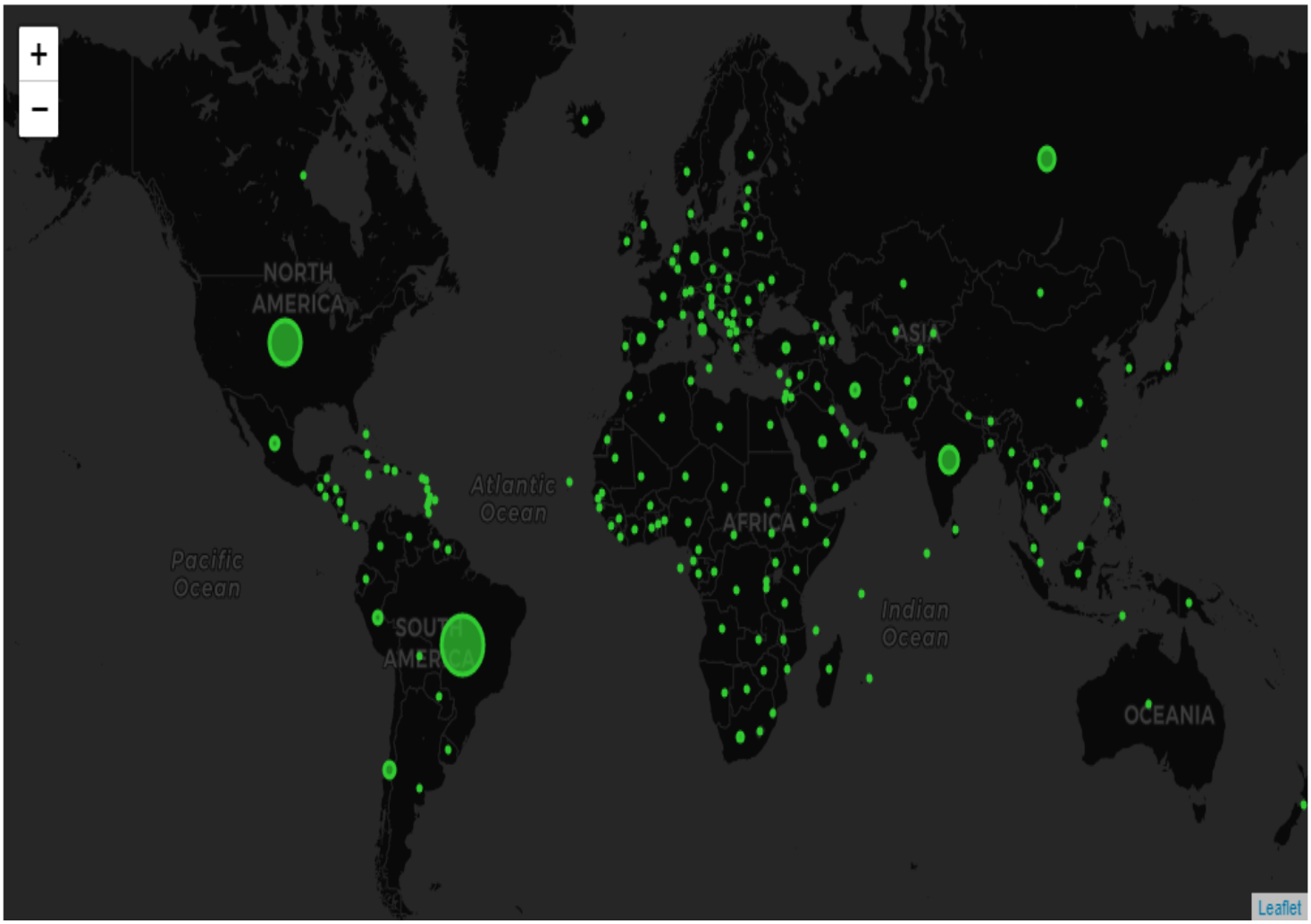
5.1. SPREAD ACROSS THE WORLD

5.1.1. NUMBER OF CONFIRMED CASES



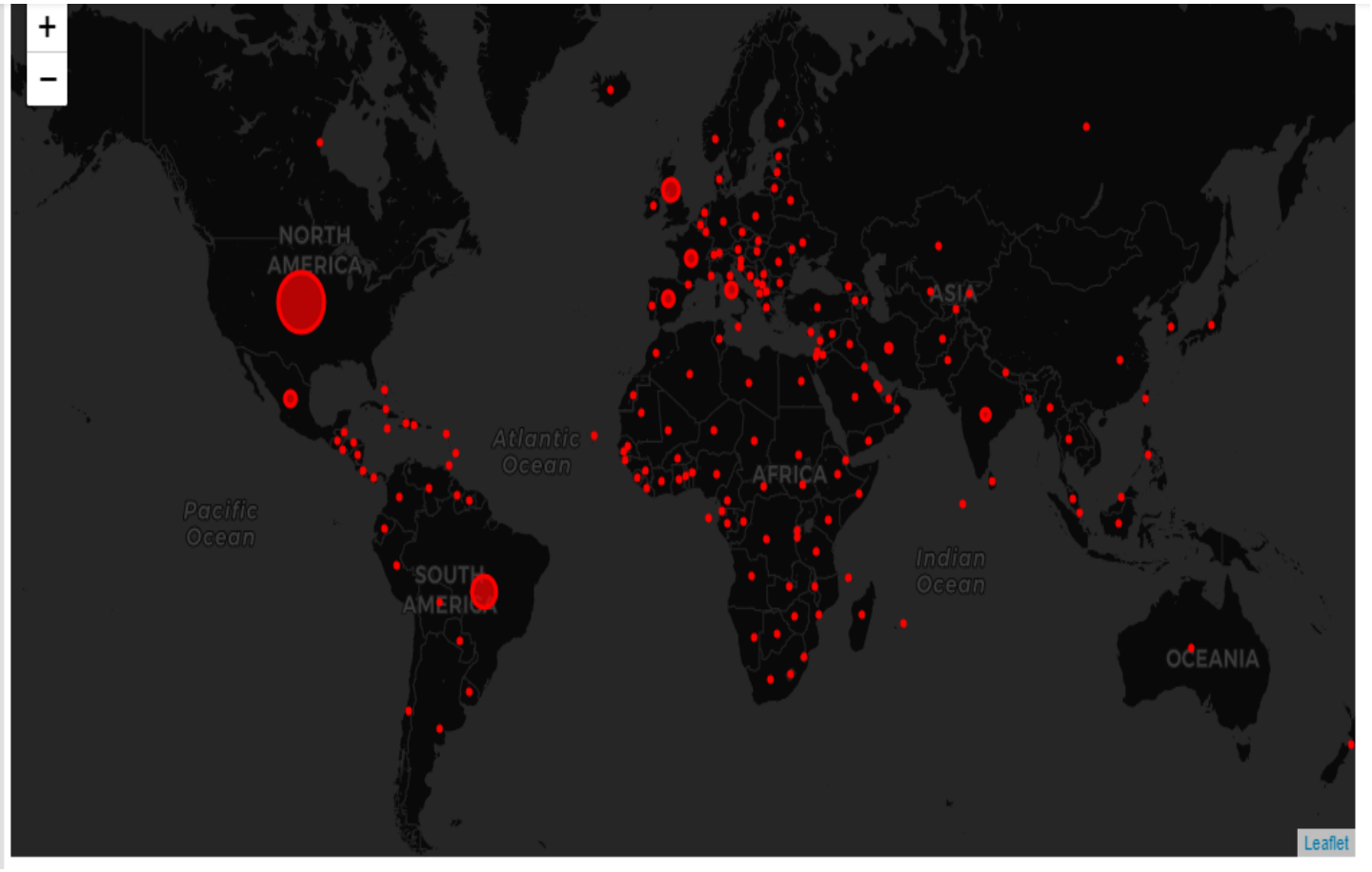
We can see that USA, Brazil,India and Russia are some of the worst hit countries by COVID-19 in terms of confirmed cases.

5.1.2. NUMBER OF RECOVERED CASES



Recovery rate in South America is higher in comparison to rate in North America. The reason could be possibly be of residents of South America having stronger immunity than that of North America.

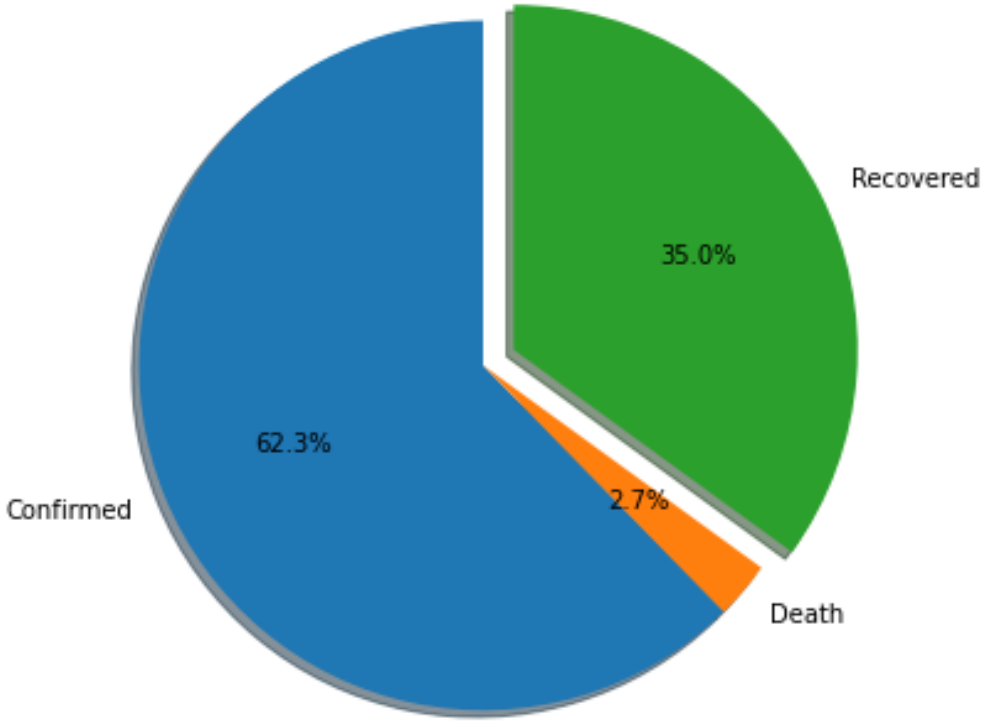
5.1.3. NUMBER OF DEATHS



Here we can see that even though India and Russia are some of the worst hit countries in terms of confirmed cases but they have managed to contain death rate.

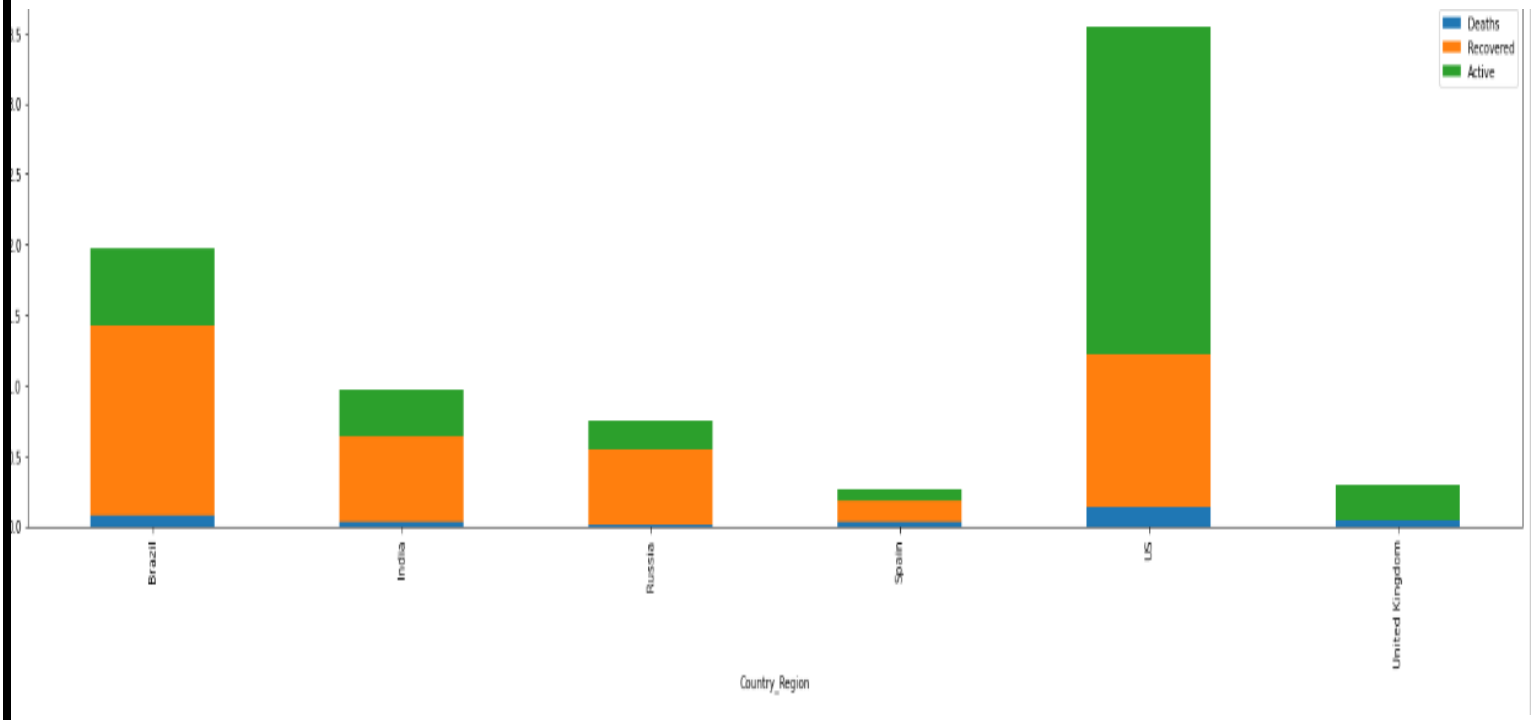
OBSERVATION: We observed that North America mainly USA has most number of confirmed cases all over the world with high number in death tolls.

5.2. PERCENTAGE OF CONFIRMED, DEATHS AND RECOVERED



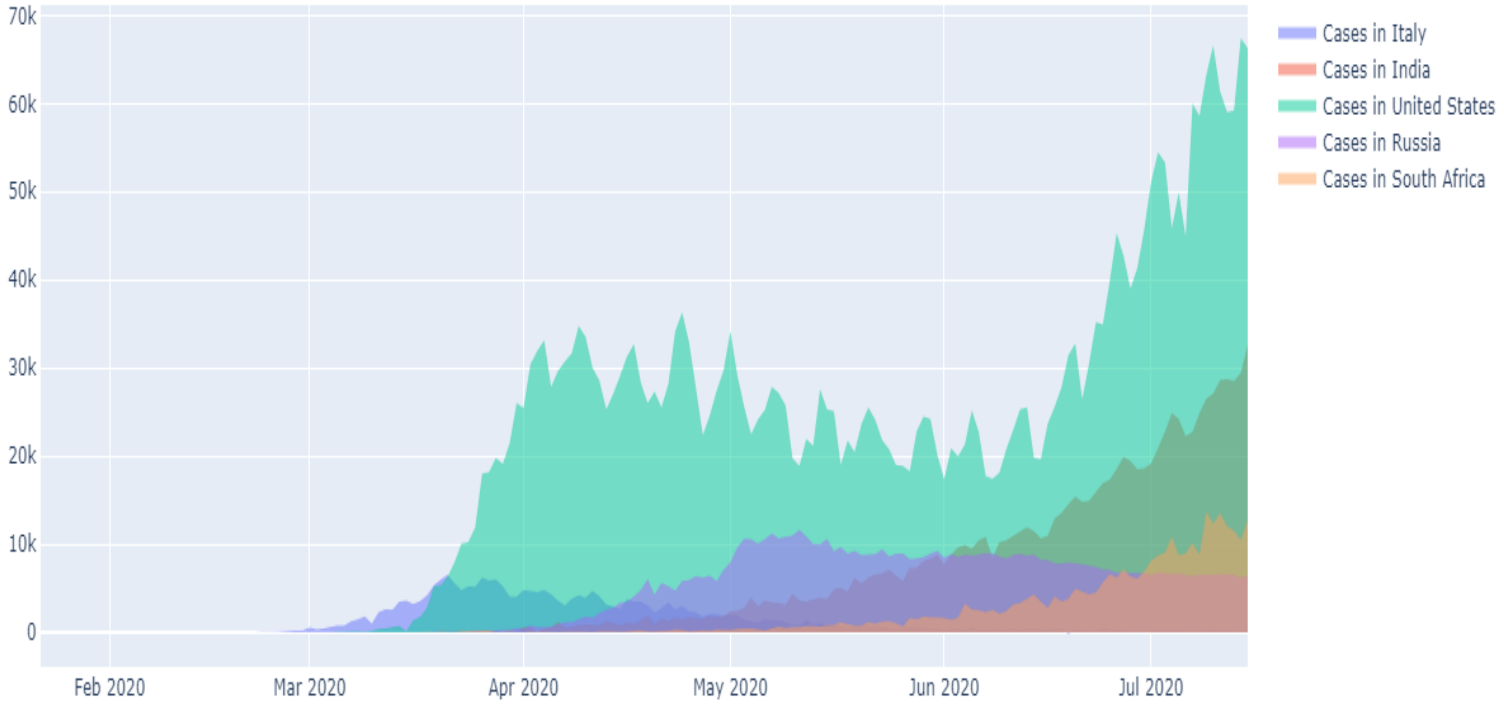
The world has managed to contain death rate to 2.7 percent while the recovery rate is 35% which means the world is slowly developing herd immunity.

5.3. TOP 6 COUNTRIES MOST AFFECTED BY CORONAVIRUS



Through **Marimekko Plot**, we understood sections of deaths, recovered and active cases for 6 top countries which have most cases, where US stands at top followed by Brazil in most affected countries.

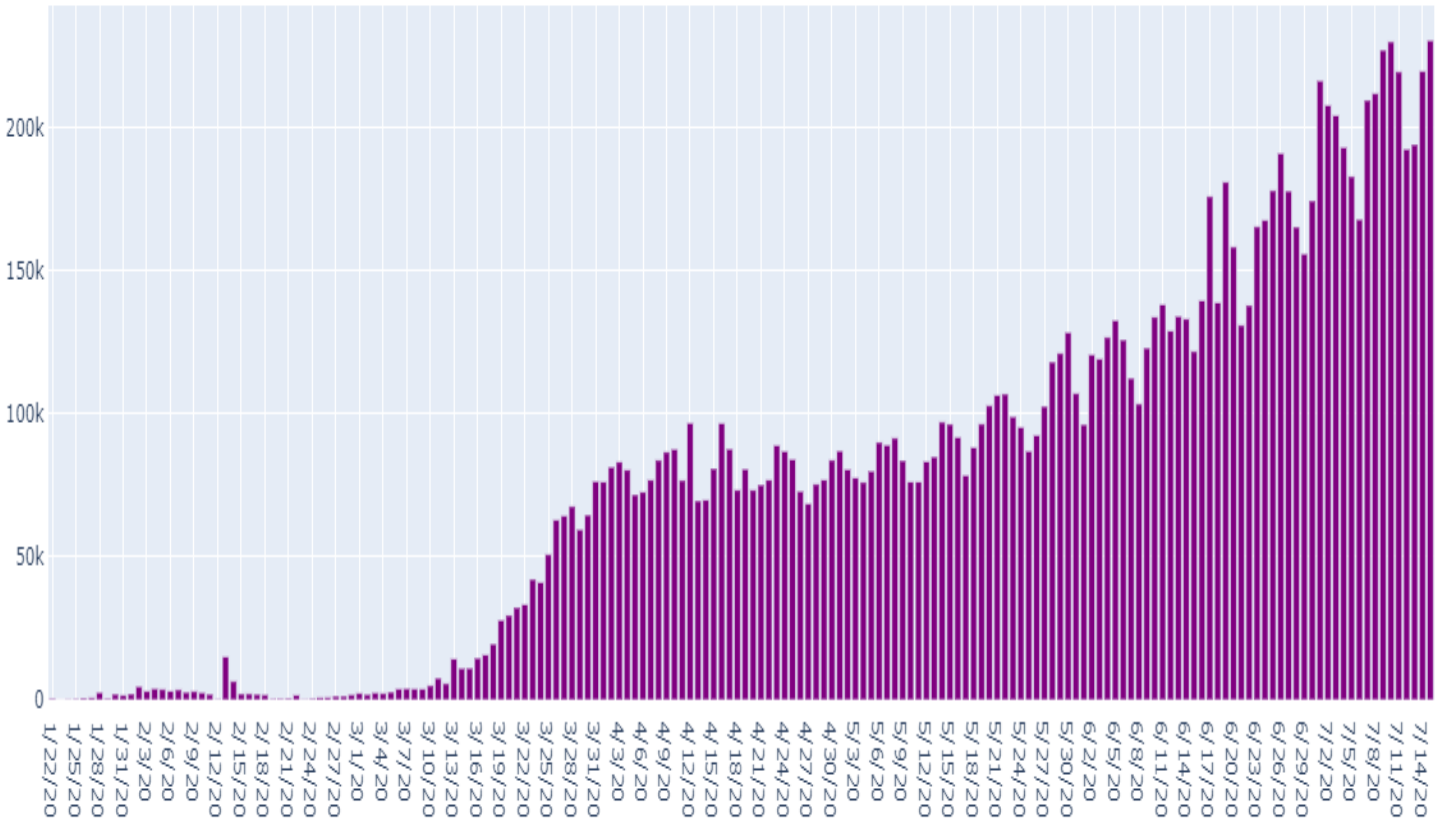
5.4. CASES RISING EVERYDAY



Here we can see that though United States and Italy were growing at same rate in April but Italy has managed to contain the number of cases in its country and has actually flattened the curve while USA continues to grow.

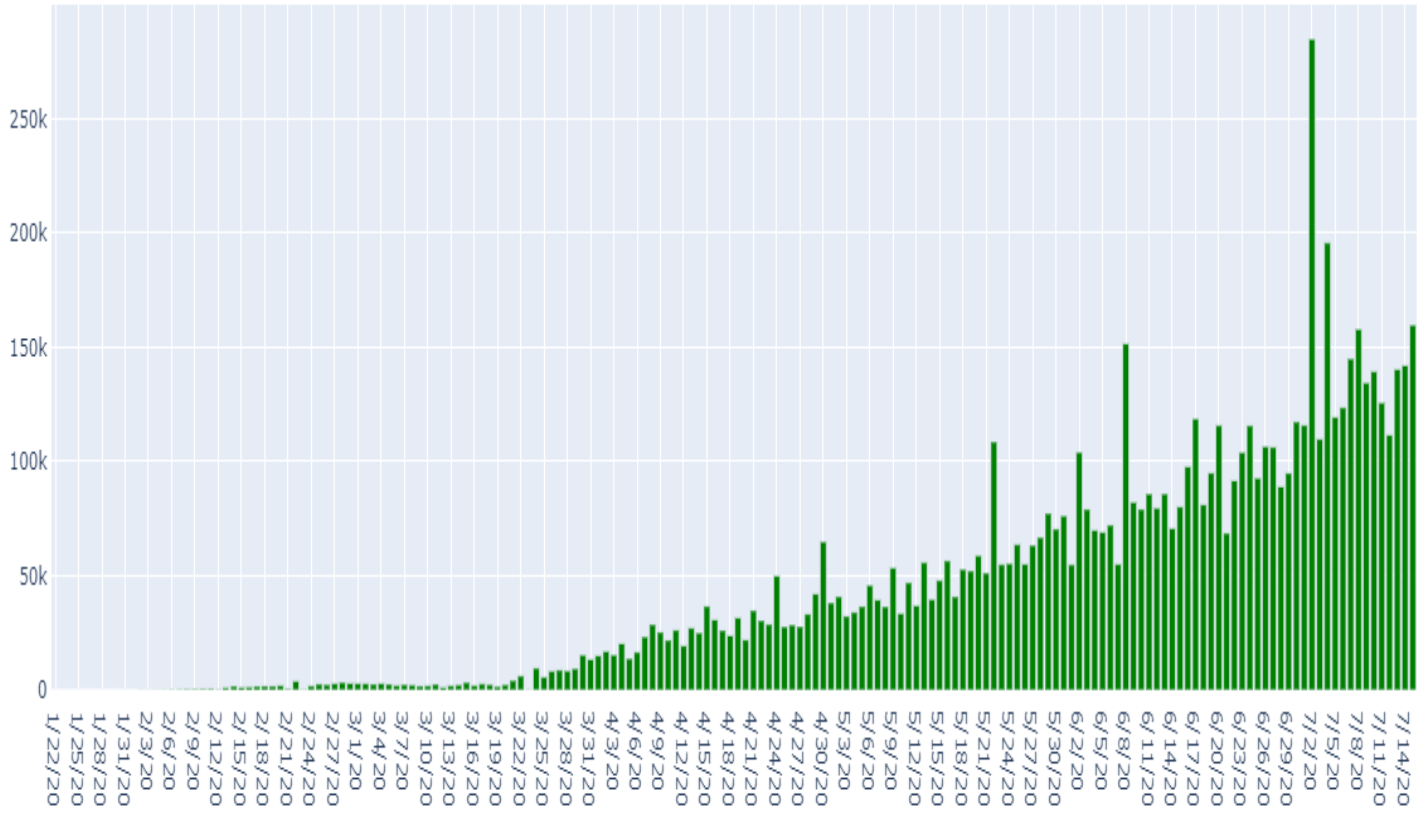
5.5. DAY WISE CASES

5.5.1 CONFIRMED CASES



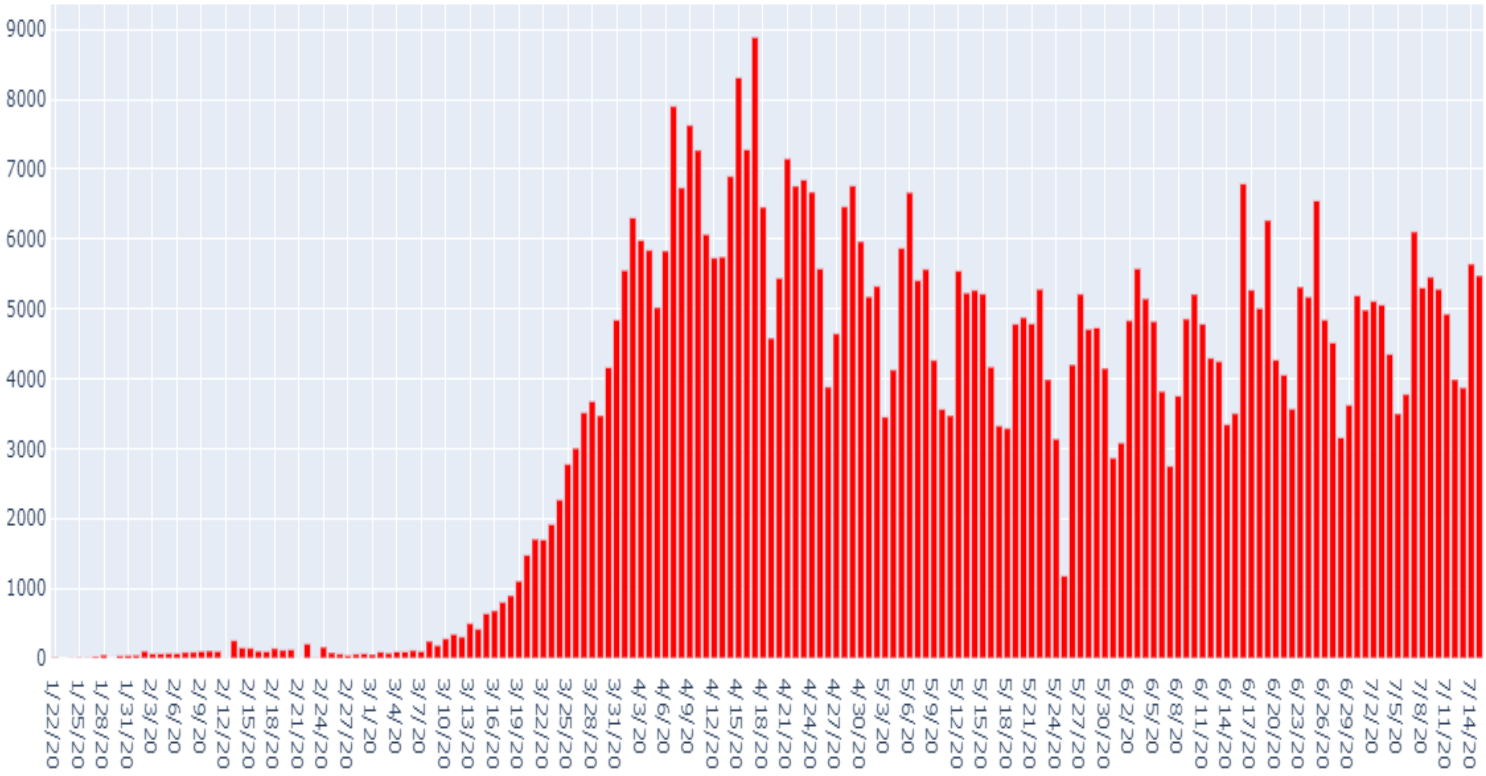
Here we can see that since many countries are coming out of the lockdown cases are increasing daily at an exponential rate.

5.5.2 DAILY RECOVERIES ALL OVER THE WORLD



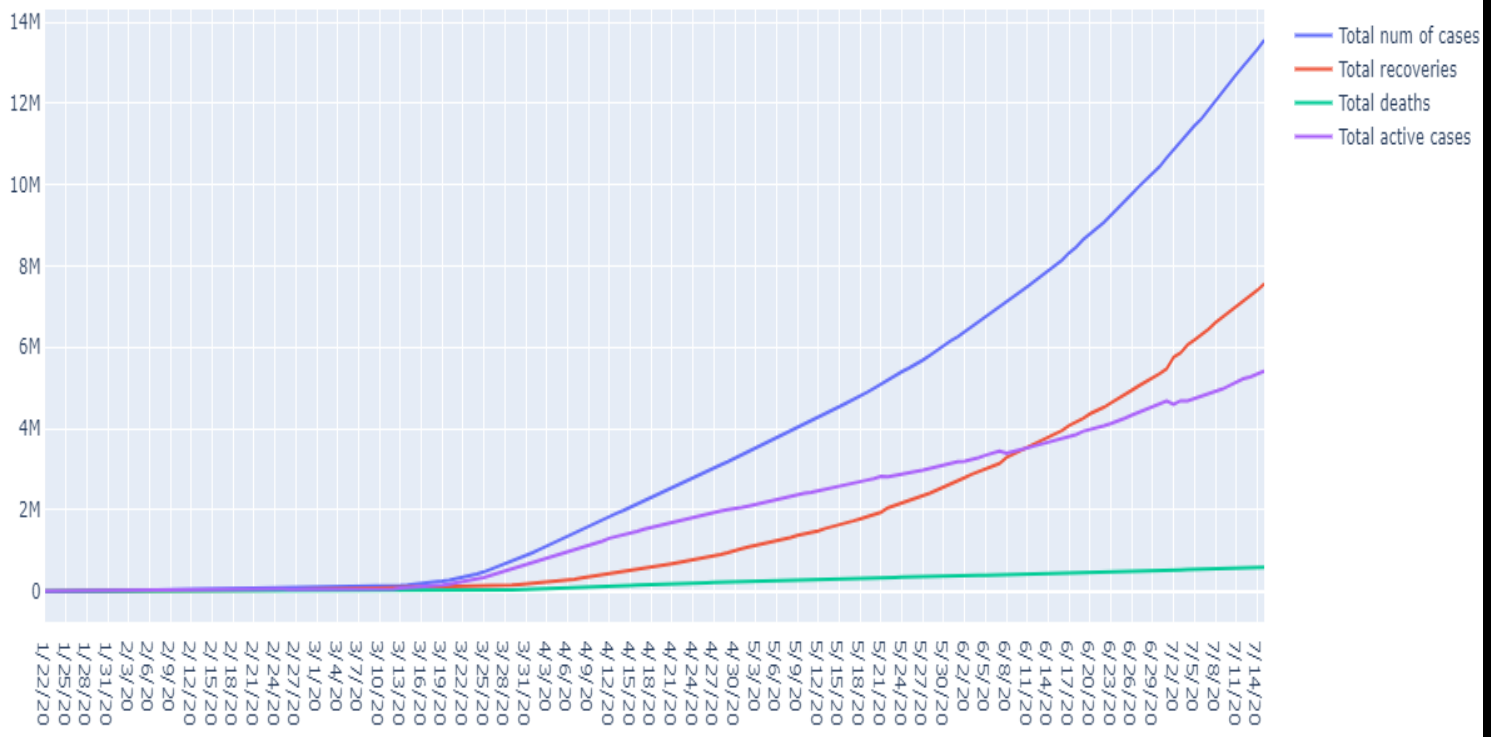
From this we can infer that since many companies are coming up with various medicines, recovery rate has increased. Since people around the globe are focusing more on their immunity we are seeing people getting recovered at a faster rate but the question remains that can we achieve herd immunity before causing a lot of deaths?

5.5.3 DAILY DEATHS ALL OVER THE WORLD



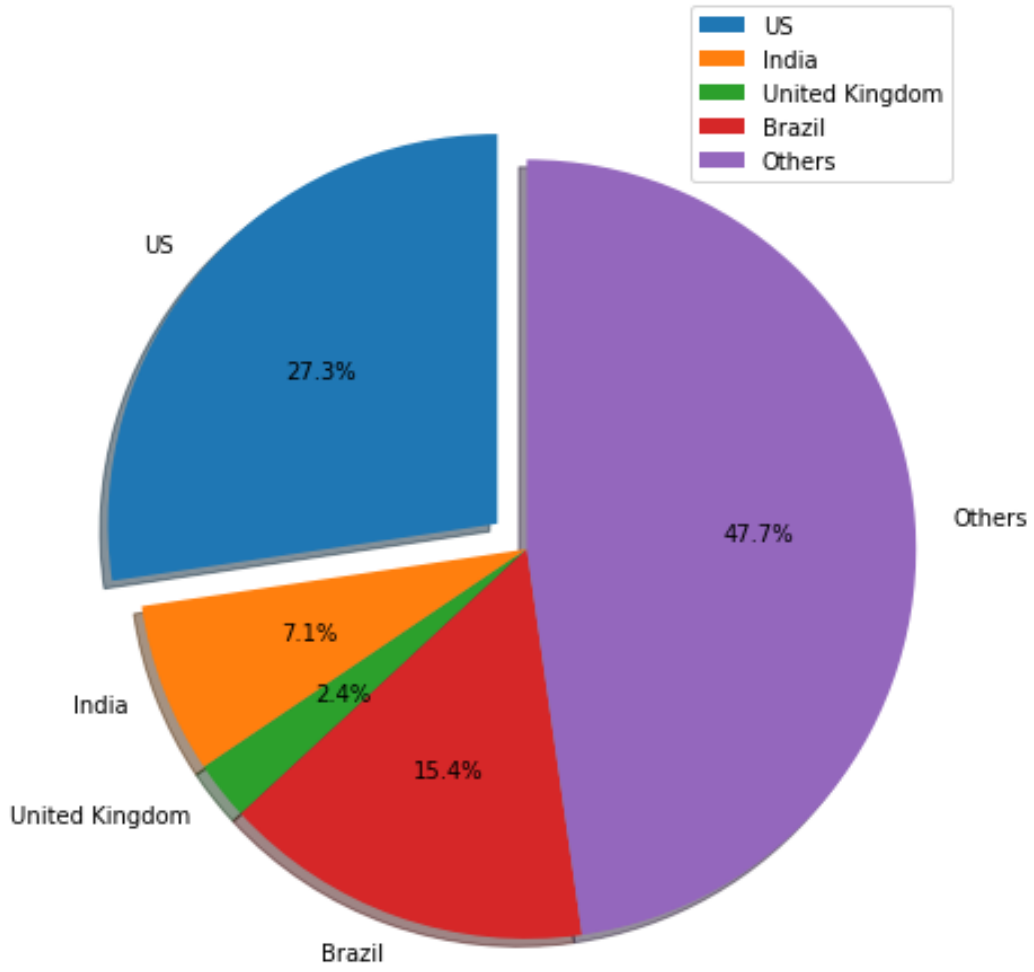
We can see that we are somehow reducing number of deaths as more and more healthcare workers are coming with more and more combination of treatments and the people around the world are taking care of their immunity.

5.6. COMPARING DAILY CONFIRMED, DEATHS, ACTIVE AND RECOVERED



We can see that though total numbers of cases are increasing daily but number of recoveries have overtaken number of active cases which is a good sign of recovery. Also active cases curve is starting to flatten out which means both our confirmed cases and recoveries are increasing exponentially.

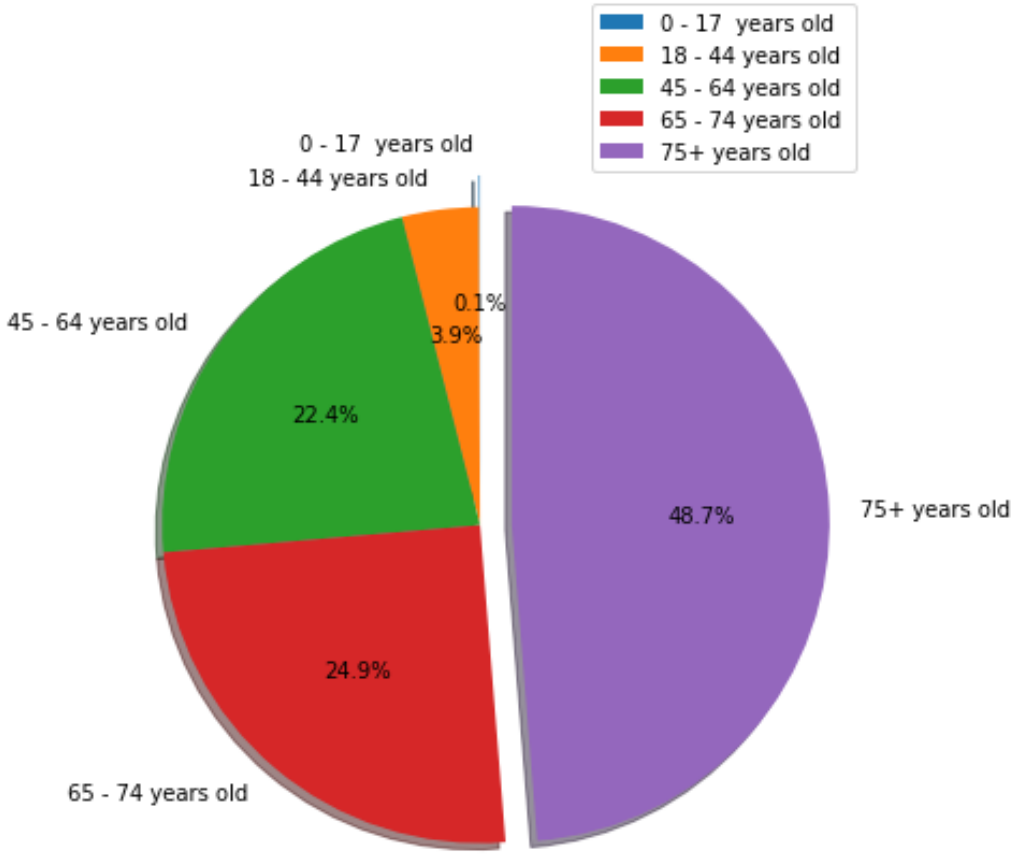
5.7. COMPARING CONFIRMED CASES USING PIE-CHART



Only 4 countries account for more than 50% of cases in the whole world.

5.8. MORTALITY RATE AS PER AGE

Here we got data from New York City Health as of May 13, 2020 of 15230 deaths of individuals and we will try to show which age group is affected the most by COVID-19. Using Beautiful Soup, we scratched the table we needed for this visualization.

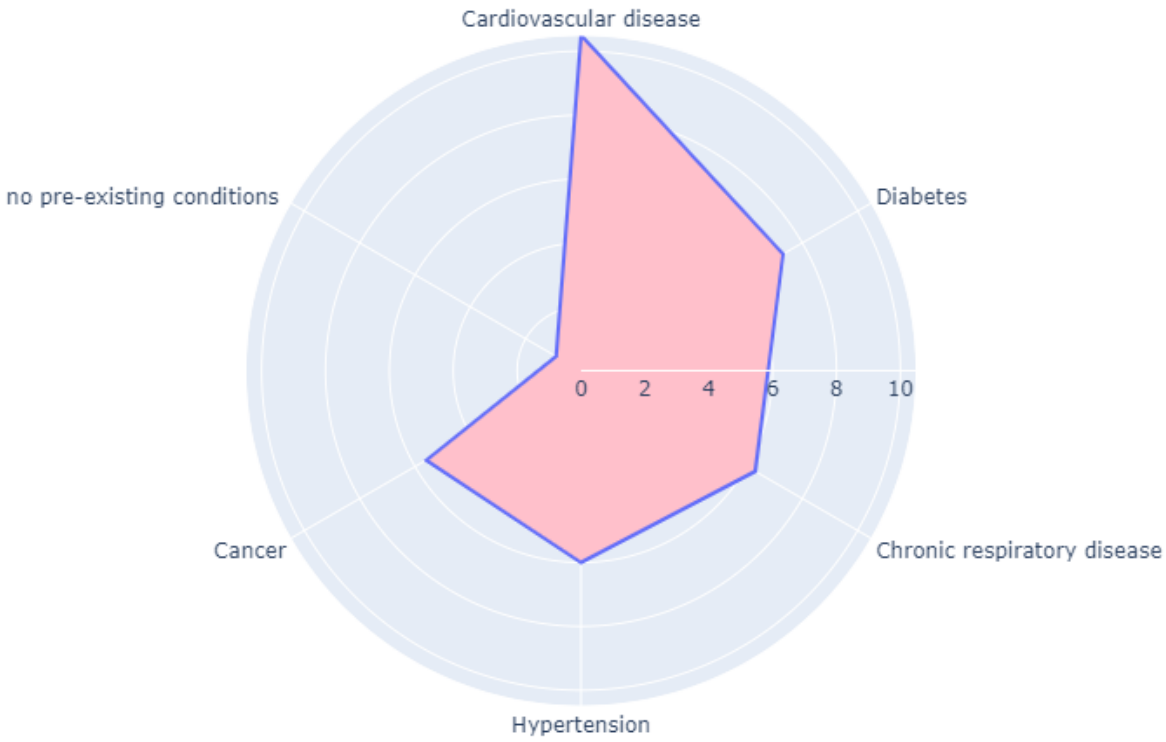


COVID-19 is hardly affecting 0-17 years old and mostly affecting 75+ years old.

5.9. DEATH RATE

Death Rate= number of deaths / number of cases = probability of dying if infected by the virus (%).

This probability differs depending on pre-existing condition. The percentage shown below does NOT represent in any way the share of deaths by pre-existing condition. Rather, it represents, for a patient with a given pre-existing condition, the risk of dying if infected by COVID-19.



People with CARDIOVASCULAR DISEASE are worst hit by COVID-19.

6. MODEL DEVELOPMENT

Model development is an *iterative* process, in which many models are derived, tested and built upon until a model fitting the desired criteria is built. A model or estimator can be thought of as a mathematical equation used to predict a value given one or more other value. Usually the more relevant your data you have, the more accurate your model is.

It is a set of tools and techniques used to understand and analyze how to collect, update, and store data. It is a critical skill with discovering, analyzing, and specifying changes to how software systems create and maintain information.

Our main aim for this project is to build a model that predicts COVID-19 cases for next 7 days. As we know, COVID-19 cases are regularly increasing, this shows an upward trend. Therefore, we need a model for future forecasting with an upward trend.

ARIMA MODEL is a popular and widely used statistical method for time series forecasting. ARIMA is an acronym that stands for Auto-Regressive Integrated Moving Average. It is a class of model that captures a suite of different standard temporal structures in time series data. It explicitly caters to a suite of standard structures in time series data, and as such provides a simple yet powerful method for making skillful time series forecasts.

What's wrong with ARIMA?

As its name suggests, it supports both an autoregressive and moving average elements. The integrated element refers to differencing allowing the method to support time series data with a trend.

A problem with ARIMA is that it does not support seasonal data. That is a time series with a repeating cycle.

ARIMA expects data that is either not seasonal or has the seasonal component removed, e.g. seasonally adjusted via methods such as seasonal differencing.

Also, ARIMA doesn't give much accurate predictions for trend in comparison to SARIMAX.

SARIMAX MODEL

Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component.

It adds three new hyper parameters to specify the auto regression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

The estimation model assumes that the current situation prevails with regard to mingling of people and recovery of infected individuals.

How to configure SARIMAX?

Configuring a SARIMA requires selecting hyper parameters for both the trend and seasonal elements of the series.

Trend Elements

There are three trend elements that require configuration.

- p: Trend auto regression order.
- d: Trend difference order.
- q: Trend moving average order.

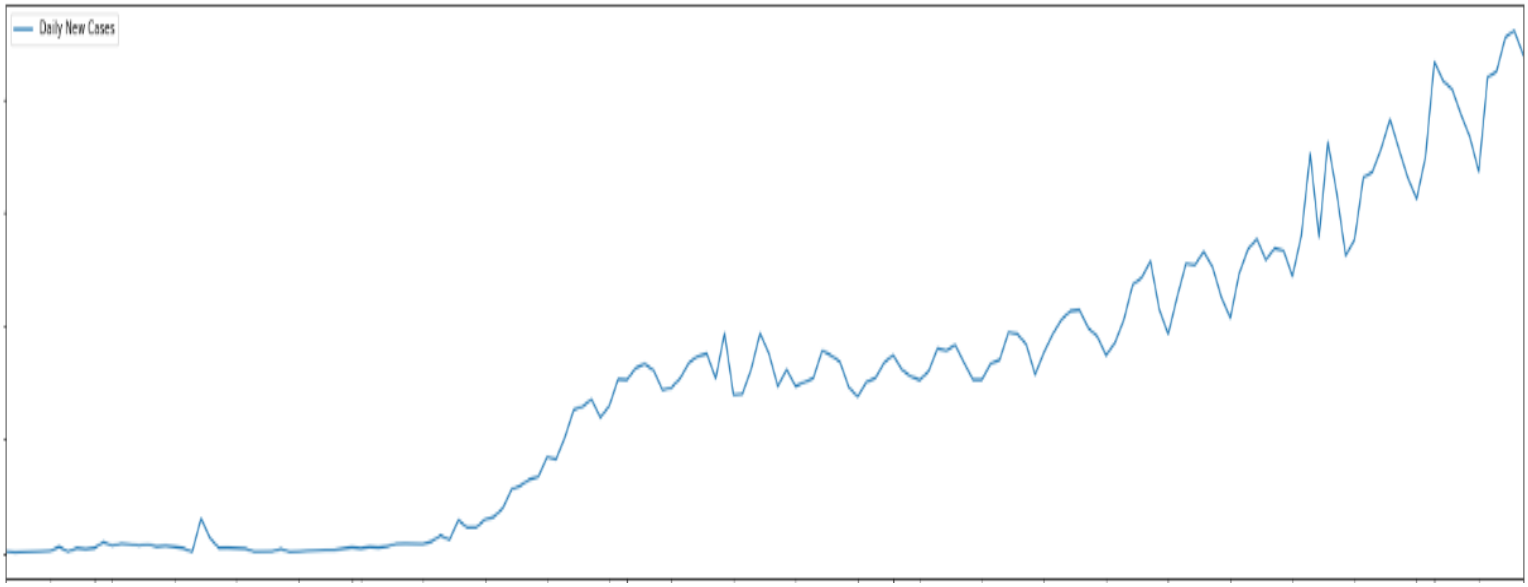
Seasonal Elements

There are four seasonal elements that are not part of ARIMA that must be configured; they are:

- **P**: Seasonal autoregressive order.
- **D**: Seasonal difference order.
- **Q**: Seasonal moving average order.
- **m**: The number of time steps for a single seasonal period.

As in our dataset, daily new cases show an upward trend we do not need a seasonal component, so we will neglect the seasonal component by keeping it as 0.

7. MODEL EVALUATION USING VISUALIZATION



How to use SARIMAX?

The SARIMAX time series forecasting method is supported in Python via the Stats models library.

To use SARIMAX there are three steps, they are:

1. Define the model.
2. Fit the defined model.
3. Make a prediction with the fit model.

8. MODEL PREDICTION

8.1. Prediction Dataset- This is the dataset on which our model will be built. One column here is dates from 22nd January to 11th July. Using this we will predict new cases for next 7 days.

dates	Daily New Cases
2020-01-22	555.0
2020-01-23	99.0
2020-01-24	287.0
2020-01-25	493.0
2020-01-26	684.0

8.2. Finding Parameters using Auto-ARIMA- The accurate values of p, d, q & m are calculated in this step.

```
Fit ARIMA: order=(2, 1, 2) seasonal_order=(0, 0, 0, 1); AIC=3664.932, BIC=3683.782, Fit time=0.476 seconds
Fit ARIMA: order=(0, 1, 0) seasonal_order=(0, 0, 0, 1); AIC=3660.537, BIC=3666.821, Fit time=0.015 seconds
Fit ARIMA: order=(1, 1, 0) seasonal_order=(0, 0, 0, 1); AIC=3659.640, BIC=3669.065, Fit time=0.025 seconds
Fit ARIMA: order=(0, 1, 1) seasonal_order=(0, 0, 0, 1); AIC=3657.990, BIC=3667.415, Fit time=0.034 seconds
Fit ARIMA: order=(1, 1, 1) seasonal_order=(0, 0, 0, 1); AIC=3659.297, BIC=3671.864, Fit time=0.093 seconds
Fit ARIMA: order=(0, 1, 2) seasonal_order=(0, 0, 0, 1); AIC=3643.169, BIC=3655.736, Fit time=0.193 seconds
Fit ARIMA: order=(1, 1, 3) seasonal_order=(0, 0, 0, 1); AIC=3642.366, BIC=3661.216, Fit time=0.330 seconds
Fit ARIMA: order=(0, 1, 3) seasonal_order=(0, 0, 0, 1); AIC=3665.731, BIC=3681.439, Fit time=0.095 seconds
Fit ARIMA: order=(2, 1, 3) seasonal_order=(0, 0, 0, 1); AIC=3636.144, BIC=3658.136, Fit time=0.514 seconds
Fit ARIMA: order=(2, 1, 4) seasonal_order=(0, 0, 0, 1); AIC=3631.404, BIC=3656.538, Fit time=0.760 seconds
/usr/local/lib/python3.6/dist-packages/statsmodels/base/model.py:512: ConvergenceWarning:

Maximum Likelihood optimization failed to converge. Check mle_retvals

Fit ARIMA: order=(3, 1, 5) seasonal_order=(0, 0, 0, 1); AIC=3579.510, BIC=3610.927, Fit time=1.122 seconds
Fit ARIMA: order=(2, 1, 5) seasonal_order=(0, 0, 0, 1); AIC=3576.621, BIC=3604.896, Fit time=0.916 seconds
Fit ARIMA: order=(1, 1, 4) seasonal_order=(0, 0, 0, 1); AIC=3649.504, BIC=3671.496, Fit time=0.274 seconds
Fit ARIMA: order=(1, 1, 5) seasonal_order=(0, 0, 0, 1); AIC=3629.207, BIC=3654.340, Fit time=0.248 seconds
Total fit time: 5.111 seconds
ARIMA(callback=None, disp=0, maxiter=50, method=None, order=(2, 1, 5),
      out_of_sample_size=0, scoring='mse', scoring_args={},
      seasonal_order=(0, 0, 0, 1), solver='lbfgs', start_params=None,
      suppress_warnings=False, transparams=True, trend='c')
```

8.3 FITTING MODEL- Taking the accurate values of p, d, q and m, we fit our model and now our model is ready for prediction.

Statespace Model Results

Dep. Variable:	Daily New Cases	**No. Observations:**	172
Model:	SARIMAX(2, 1, 5)	**Log Likelihood**	-1781.493
Date:	Sun, 12 Jul 2020	**AIC**	3578.985
Time:	16:20:40	**BIC**	3604.118
Sample:	01-22-2020 - 07-11-2020	**HQIC**	3589.183
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.2363	0.023	53.980	0.000	1.191	1.281
ar.L2	-0.9933	0.016	-60.375	0.000	-1.026	-0.961
ma.L1	-1.8527	0.067	-27.553	0.000	-1.984	-1.721
ma.L2	1.7562	0.149	11.784	0.000	1.464	2.048
ma.L3	-0.4126	0.220	-1.875	0.061	-0.844	0.019
ma.L4	-0.3224	0.183	-1.765	0.078	-0.680	0.036
ma.L5	0.3163	0.086	3.688	0.000	0.148	0.484
sigma2	8.753e+07	1.46e-09	5.99e+16	0.000	8.75e+07	8.75e+07

Ljung-Box (Q):	56.59	**Jarque-Bera (JB):**	87.86
Prob(Q):	0.04	**Prob(JB):**	0.00

8.4 MAKING PREDICTIONS- In this, we will use predict method to predict new cases of COVID-19 for next 7 days.

	ARIMA Predictions
2020-07-12	198567.549854
2020-07-13	198410.351769
2020-07-14	214408.369760
2020-07-15	232664.548535
2020-07-16	242981.082866
2020-07-17	237602.421820
2020-07-18	220705.446901

9. ACCURACY OF MODEL

Accuracy is defined as how close a measurement is to the true or accepted value. Accuracy is the prime metric to compare models. The accuracy tells that overall how often the model is making a correct prediction. Accuracy is important for the acceptable certainty of results obtained from point of view of expected consequences and theory targets. The inaccuracy of predicted output values is termed the *error* of the model.

ERROR RATE tells that overall how often the model is making a wrong prediction. Also, Error = 1-Accuracy.

Error rate is calculated by below formula:

$$\delta = \left| \frac{v_A - v_E}{v_E} \right| \cdot 100\%$$

δ = percent error

v_A = actual value observed

v_E = expected value

DAY WISE PREDICTION ERROR

DATE	ERROR RATE
12/07/2020	0.280%
13/07/2020	0.893%
14/07/2020	2.640%
15/07/2020	1.244%
16/07/2020	2.673%
17/07/2020	1.383%
18/07/2020	2.780%

AVERAGE DAY WISE ERROR RATE: 1.694%

10. OBSERVATIONS & CONCLUSION

From this project, we observed that:

- COVID-19 cases are increasing rapidly all over the world with U.S.A. and Brazil on top.
- Death rate is dropped to 2.7 percent all over the world.
- People with cardiovascular disease are worst affected by this disease.
- 75+ age people are hit the most by percentage of 47.7% and least 0.1% for 0-17 years old.
- Only 4 countries in the world namely United States of America, Brazil, India and United Kingdom account for more than 50% cases in the whole world.
- Death Rate is decreasing and Recovery rate is increasing which is a positive side for the people living.

Conclusion derived from this project is that there will be an increasing graph showing trend in number of cases till vaccine is developed. COVID-19 outbreak will have to wait till a vaccine is developed.

PROJECT URL:
https://colab.research.google.com/drive/1uRgXdUmR_Ru5S4mRYa-IBhU_TVioCW4H?usp=sharing