COVID-19 DATA SET ANALYSIS

CONTENTS

1. Introduction
2. About Our Dataset
3. Libraries Used
4. Graphical Representation and Analysis
5. Conclusion
6. References

Introduction

Coronaviruses are a large family of viruses causing illness in animals or humans. The outbreak first reported in Wuhan, China is spreading over the world with astounding speed and has severe consequences. The cases increased rapidly with major outbreaks in the United States, Italy, Iran, and more than 50 other countries with total cases crossing over 15 lakhs worldwide in a span of just 2-3 months.

As this epidemic continues to spread uncontrolled around the world, shops and restaurants have closed their doors, information workers have moved home, other businesses have shut down entirely, and people are social distancing and self-isolating to "flatten the curve." Many people have lost their jobs, education is affected, the economy of the country is impacted, etc.

Through this project, we aim at analyzing Coronavirus (COVID-19) spread with the help of data science and data analytics in python code. This analysis will help us to find the basis behind common notions about the virus spread from purely a dataset perspective and will help us in tackling the problem and make decisions based on the data, especially in India.

About Our Dataset

The data on COVID-19 helps us understand how the pandemic is progressing so as to take counter measures to curb the disease spread. Our dataset consists of data from 4 different sources giving us crucial information about country to country and state to state based on the total confirmed COVID-19 cases, the recovered cases and the death cases since 22nd January, 2020.The parameters of the dataset are as follows:

The **number of confirmed cases** informs us about the development of the pandemic in each country, in each state. This will tell us how the virus is affecting different regions and based on which measures can be adopted by each country. Also it helps citizens get more aware of the situation and take precautions from their side understanding the gravity of the situation.

The **number of recovered cases** informs us about the healthcare sector and facilities in each country, in each state. This will tell us how the country is reacting to the virus and how advanced is the country’s healthcare facility, how effectively the country is responding to it.

The **number of death cases** informs us how the healthcare is failing or how widely the pandemic is affecting the citizens based on their age, gender and many other factors as well.

Based on these parameters ,the analysis from the dataset will help us understand what methods to adopt, what precautions to take, which methodology to take to develop vaccination and to put an end to the virus in order to save the world from catastrophe.

Libraries Used

Numpy : Used for adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

Pandas : Used for data manipulation and analysis.

Matplotlib.pyplot : Collection of command style functions that make matplotlib work like MATLAB.

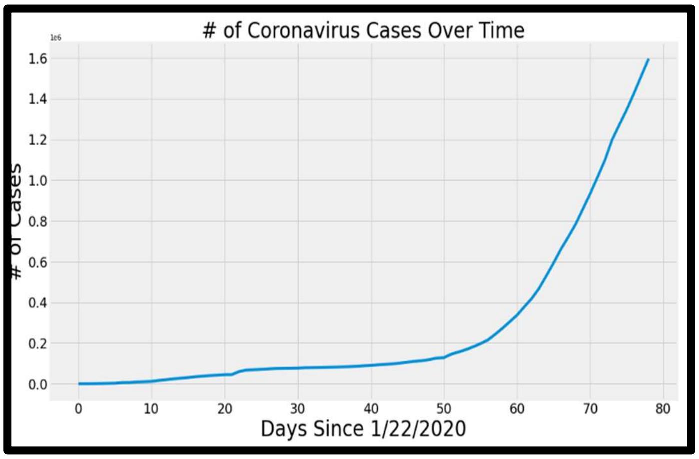
Seaborn : It is a Python data visualization library based on matplotlib.

Scikit-learn: sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.

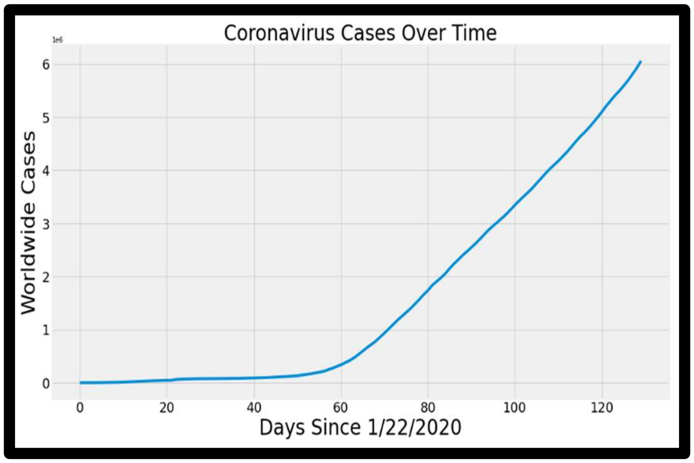
Data Interpretation and Analysis

For this project, we have only done data interpretation as part of the mid-term submission, which was done in March,2020. To better analyze and compare, we have tried to compare the latest curve (as of May,2020) with the mid term submission interpretations.

WORLD

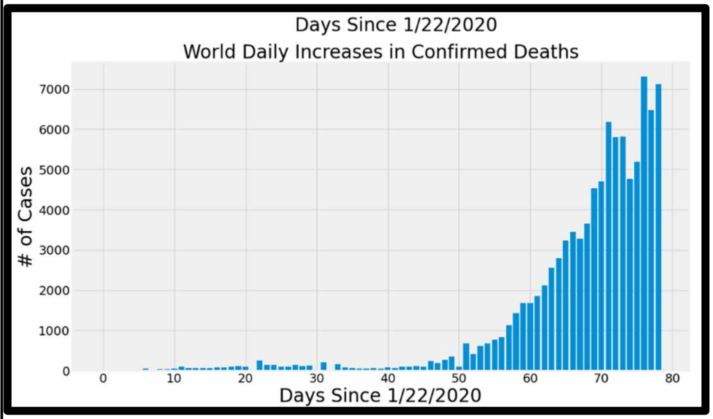


March,2020

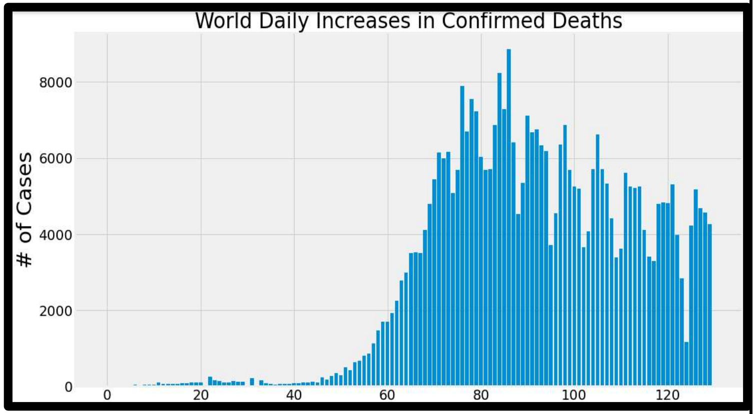


May,2020

The total cases have reached a staggering 1.5 million and continue to increase everyday by a hefty margin. As of now the recovery rate is close to 22.5 % and new cases per day continue to be around 100k. On the other hand, due to the shortage of testing kits, there might be an underreporting of cases and therefore, the total cases cannot be predicted.



March,2020

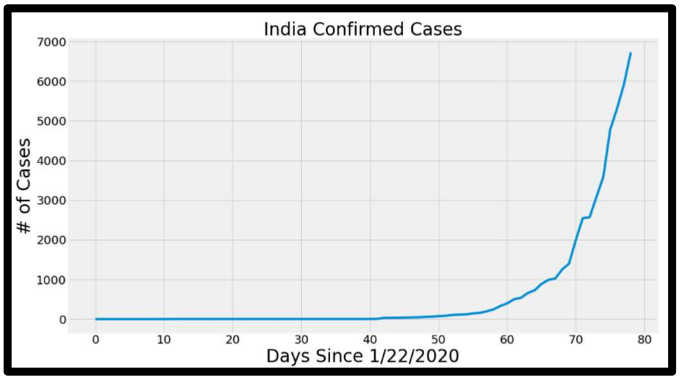


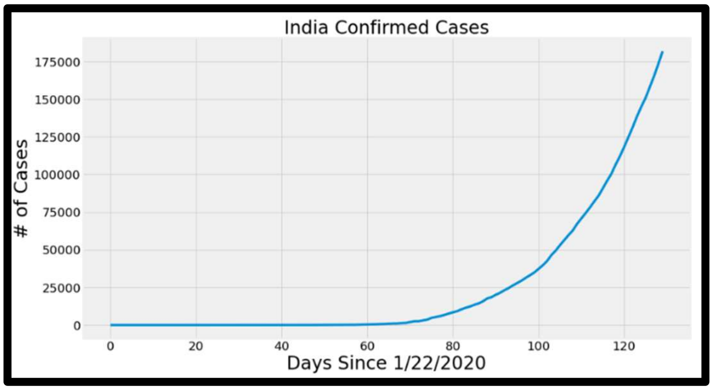
May,2020

The total deaths are on the verge of reaching to 100k. The overall mortality rate is 6% approximately and the daily death cases reached its peak 7,383. Also, the total deaths are expected to flatten the graph soon.

INDIA

Confirmed Cases



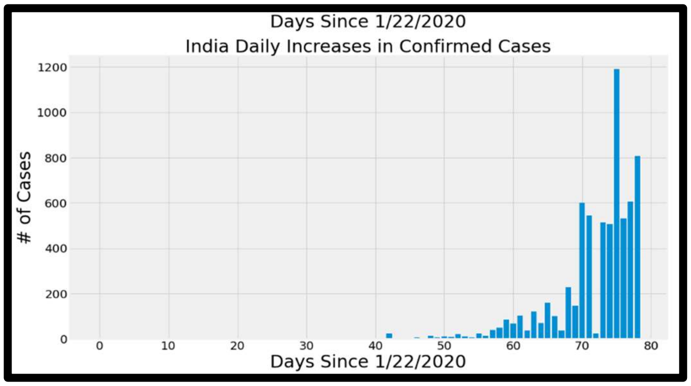


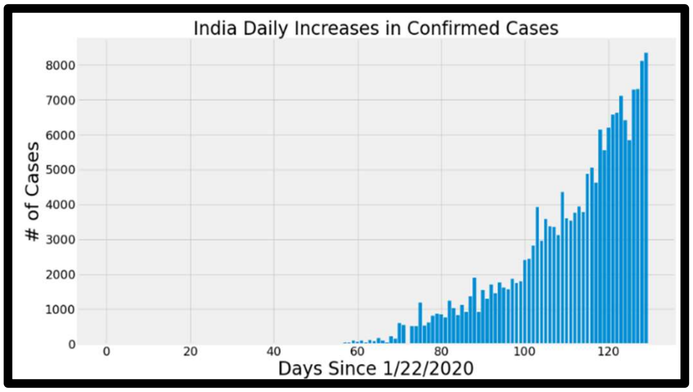
March,2020

May,2020

Inference: The graph shows us how India's rise in cases is mimicking the exponential rise seen in the early days of some Western countries. It gives us a clear view of how hard this pandemic is going to hit us. In such a densely populated country, there are very few means to control social distancing, thus seeing the exponential rise in cases. Furthermore, India has had a notable shortage of testing kits, which is undoubtedly causing a substantial underreporting of confirmed cases.

Daily Increase in Confirmed Cases



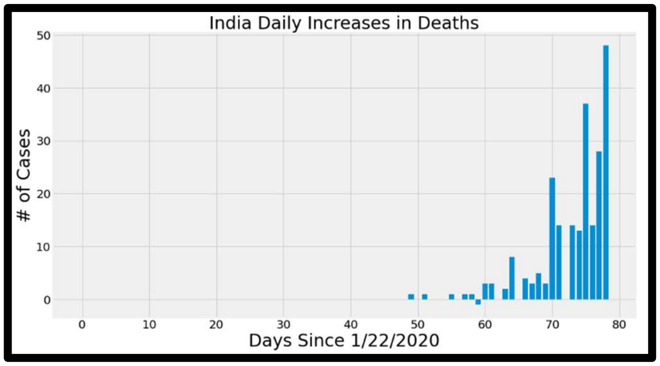


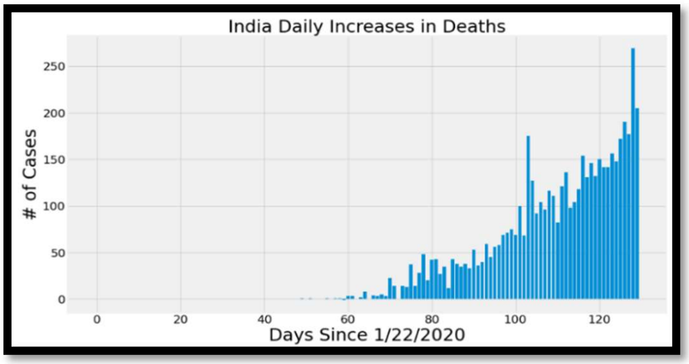
March,2020

May,2020

Inference: India has started restricting businesses, closing schools, and promoting social distancing a bit earlier than the Western Countries relatively and has issued a complete lockdown in the country. As a result of this, after achieving the peak on day 77 the number of cases in a day dropped extensively for a certain period and then started rising again (due to negligence of people). Various states will peak at different times. Currently, Maharashtra appears to have passed its peak while Tamil Nadu and Delhi are near their peaks.

Death Cases



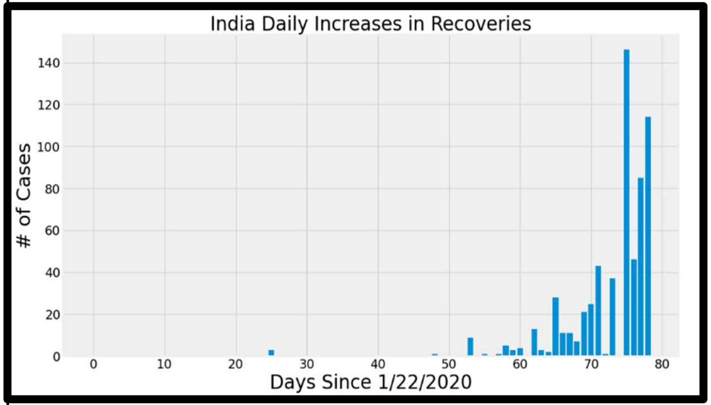


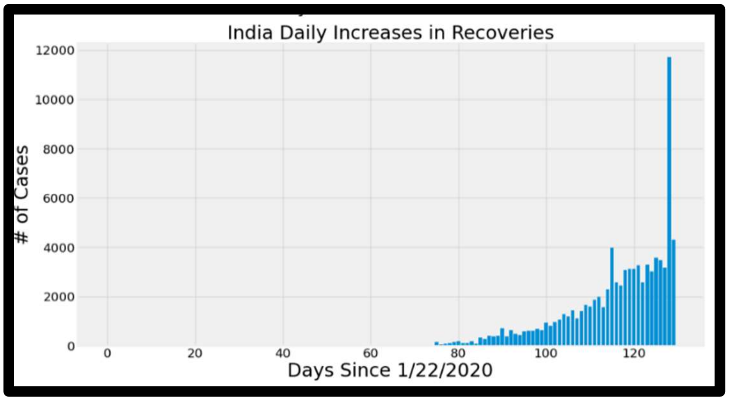
March,2020

May,2020

Inference: India's peak day for the deaths was on day 80. The lag between India's peaks for daily confirmed cases and daily deaths is 3 days. However, both the daily number of new confirmed cases and daily increase in deaths has started to rise a bit recently. Currently the mortality rate in India is close to 3%.

Recovery Cases



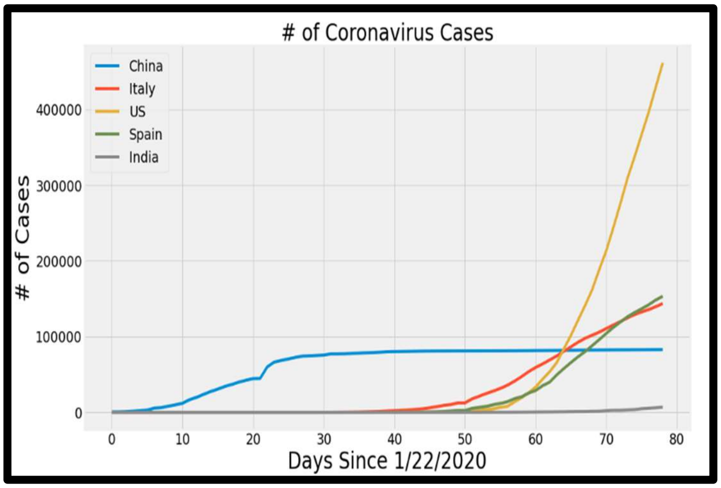


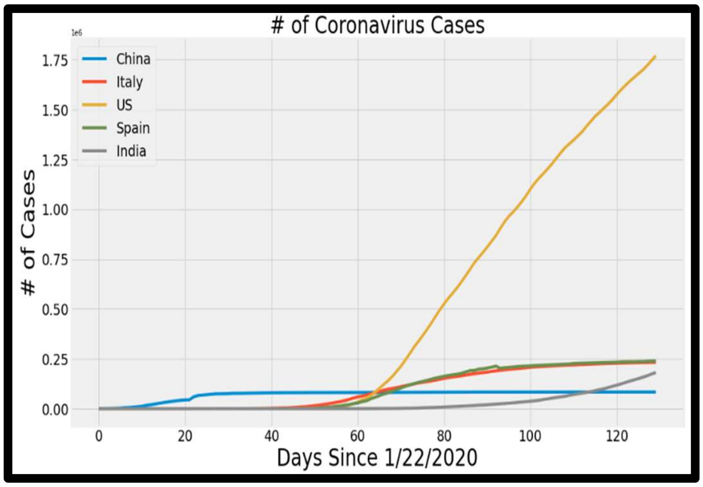
March,2020

May,2020

Inference: The total recoveries in India is approximately thrice the total deaths. The peak for this graph is the same as the peak for the Daily Confirmed cases i.e., on Day 77. The lag between India's peaks for daily death cases and daily recoveries is close to 100, which is a good sign.

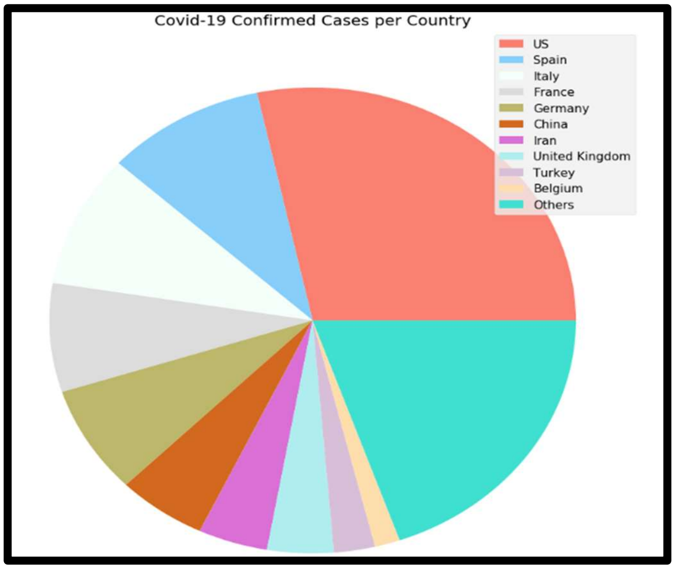
COUNTRY TO COUNTRY

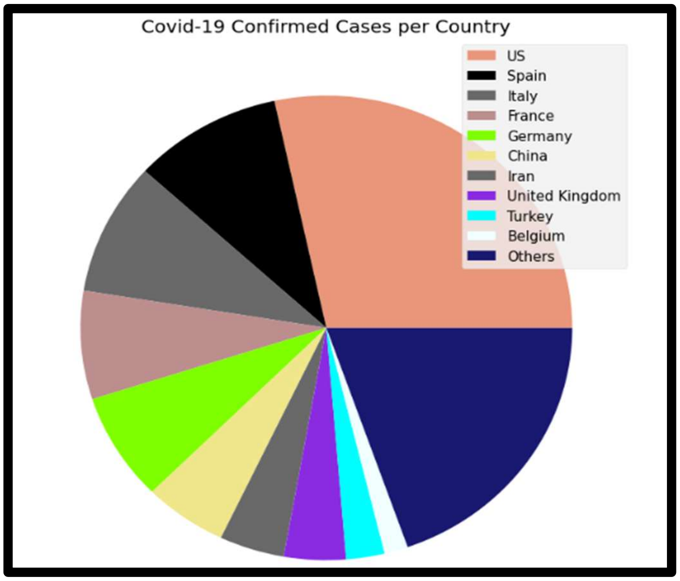




March,2020

May,2020





Currently, USA leads the chart with close to half a million cases, with Spain, France, Italy and Germany lagging just behind them with around 150k cases each.

References

1. https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv
2. https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_global.csv
3. https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_recovered_global.csv
4. https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/04-08-2020.csv

Henceforth we conclude our analysis here furthermore we will be moving towards the prediction section of the project.