**IBM Capstone Project – Covid19 Data Analysis**

By Debesh Roy

10-Sep-2020

1. Introduction

As Covid19 is the one of the main discussion topics across Globe, wanted to explore the World data and see how the graph is moving, either upward or downward.

Well, probably all these are already available over net, but I just wanted to go out of the common trend to follow the same kind of dataset to submit the project report.

1. Background

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus.

Most people who fall sick with COVID-19 will experience mild to moderate symptoms and recover without special treatment.

The virus that causes COVID-19 is mainly transmitted through droplets generated when an infected person coughs, sneezes, or exhales. These droplets are too heavy to hang in the air, and quickly fall on floors or surfaces. One can be infected by breathing in the virus if s/he is within close proximity of someone who has COVID-19, or by touching a contaminated surface and then own eyes, nose or mouth.

In this project we will try to find the trend of the various data, like World, Highest countries, digging down further within a country and identify the county level trend.

We will use the data science mechanism to generate more information about the nearby medical venues, which will help interested/impacted people to visit for consultation or medical assistance.

1. Data
   1. Data Sources

The related information about Covid19 got downloaded from multiple websites, mentioned below:

World Master: <https://covid.ourworldindata.org/data/ecdc/full_data.csv>

USA State Master: <https://covidtracking.com/api/v1/states/daily.csv>

The above USA file is having state code, this needs to be joined to get the full state names.

US State Code & Name: <https://worldpopulationreview.com/states/state-abbreviations>'

US State & County Master: <https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv>

We also downloaded the World Countries json and US States json from <https://github.com>

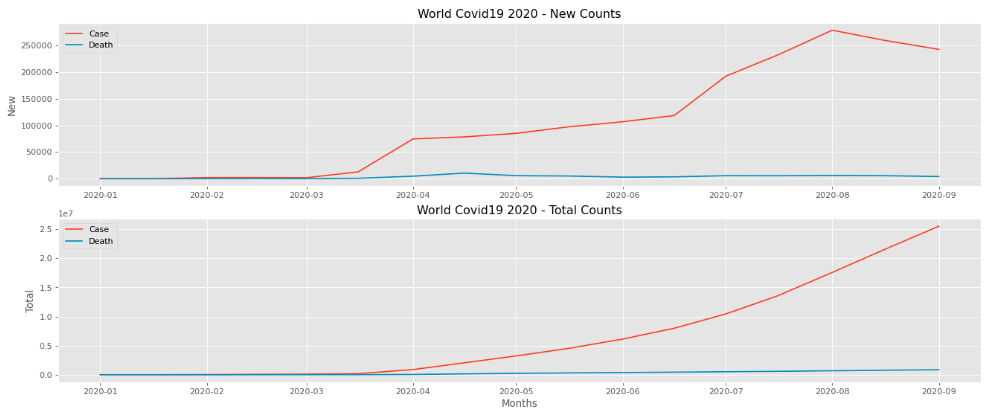
* 1. Data Extraction, Transformation, Loading

As usual, Data Cleansing is one of the most important work for Data Science, and Python is one of the leading tools to do so, Apart from Python code, libraries like Pandas, Numpy are just few to name.

Downloading the raw data from net, extracting the required data along with sorting and indexing is the process to shape and tune the data.

In ETL, Transformation is the backbone of data analysis, and mainly Pandas DataFrame is used for the same.

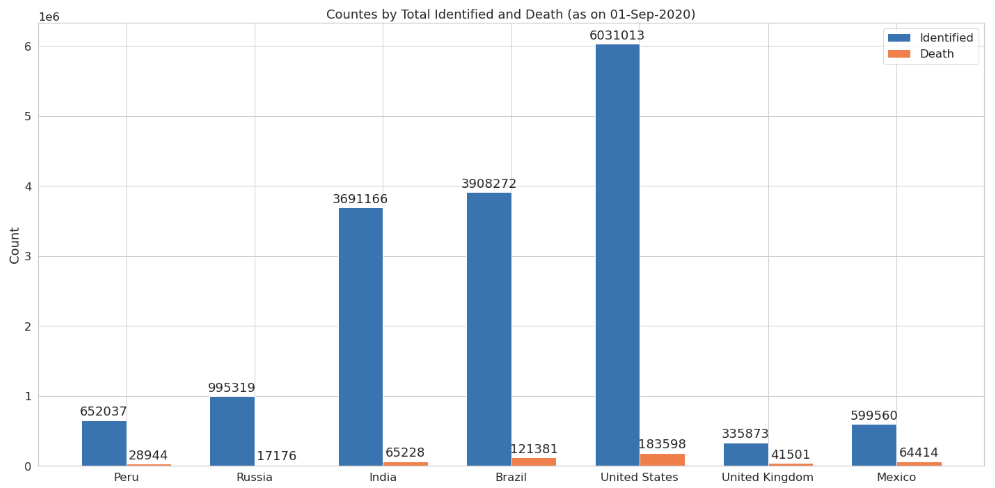
1. Representation and Analysis
   1. We compared the world data, how it increased month over month, both new identified cases and total identified cases.

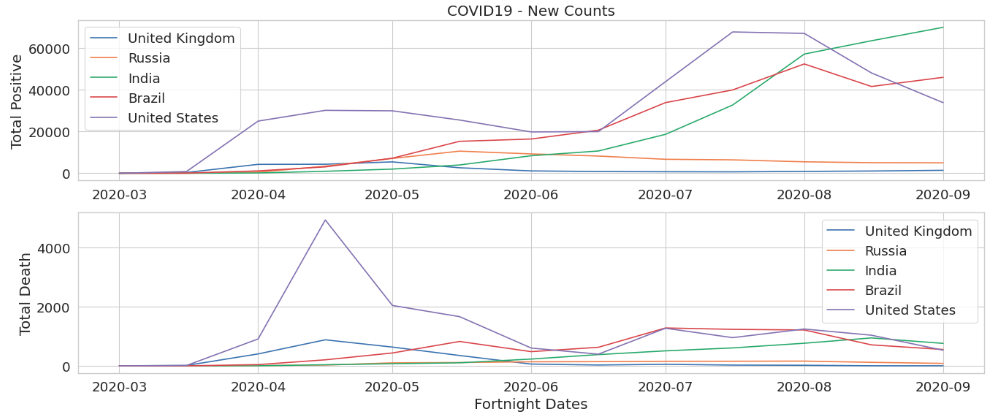


Using Python Seaborn chart, we also compared how the total death increased along with the total identified cases.

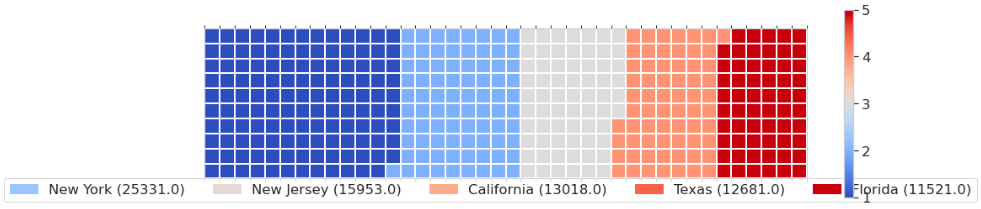
* 1. Country specific comparison

We used the bar graph to compare the leading countries identified and death counts





* 1. Waffle used to visualize the death between USA states (high)



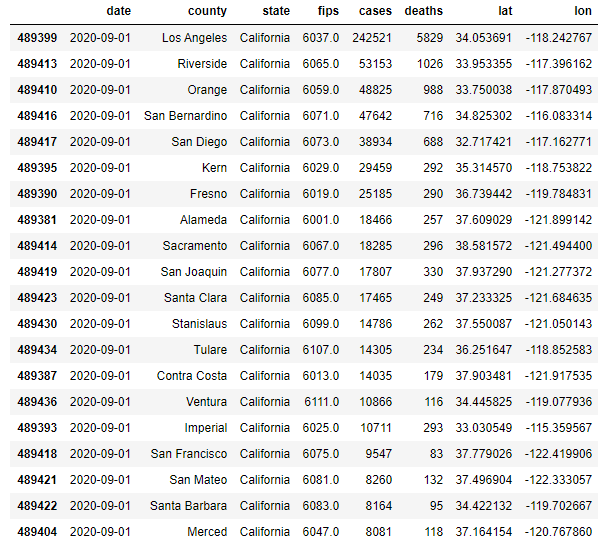
1. Methodology

The base of the project is Python, where we extensively used multiple important libraries, i.e. pandas, numpy, geocoder, folium, and many more.

Used pandas’ “read\_html” to read the information from wiki URL. Used DataFrame feature extensively to handle data, i.e. transforming, cleansing, etc.

This data handling helps to identify the State & County wise highest counts and the growth rate.

Used Florium map feature, which helped stakeholder to visualize the surrounding venues, categories, neighborhoods, etc.





1. Results

Using the Python and corresponding libraries, we identified highest countries, either by total identified positive cases, or by total reported deaths.

This summarized data can help management to put the governance in place to control the pandemic.

6. Discussion

Well, the purpose of this Python notebook is to identify the risky or high risky zones based on the identified positive cases, or total death. It also can be utilized to monitor the improvement.

Along with the data point, from the Florium map one can also identify the near around venues and approx. distances.

7. Conclusion

In this study, I tries to analyze the relationship between different countries, how different numbers are relating with each other. After going through the detailed analysis between countries, states, counties, once can easily get a fair knowledge on the surroundings and how this pandemic is revolving across regions. This study mainly focused on different features, stats to foresee the trend.