Data Warehousing and Business Intelligence

Apratim Mishra

Covid 19 Tracking and Insights

Objective

The main aim of this project is to present an exploratory analysis of the global covid-19 or coronavirus impact in terms of cases, recovery, deaths, and other important factors.

Coronavirus area a large family of viruses which may cause illness in humans and animals. In humans, several coronaviruses are known to cause respiratory infections from common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS).

This study first looks at coronavirus at a global scale and then tracks down changes in 3 countries: United States of America , India, and South Korea.

Data

The data is obtained from Kaggle, however, uploaded data on the website comes from several sources including the Johns Hopkins Repository. [1][2][3]

The first source of the data is the Covid-19 tracking project and the NY Times.[4][5]

Method

The initial step included gathering relevant data and preprocessing it to a presentable format. Preprocessing is done with Python in a Jupyter Notebook which is also submitted in the assignment. This includes creating and modifying columns to be ultimately be uploaded to the Power BI software. The csv files generated are added to the submission

In the assignment, I use Power Bi to present important reports which includes finer details, and a final dashboard is also presented which presents an overarching theme of the virus across the globe. This presents:

- Trends of the virus across time
- Trends across geographic regions
- Tracking testing measures and comparison of states and provinces within a country
- 3 countries: United States of America, India, and South Korea are selected for finer details as additional data was available from them. The time period for this data is different from the global comparison based on data limitations.

[1] https://www.kaggle.com/kimjihoo/coronavirusdataset?select=Region.csv
[2] https://www.kaggle.com/sudalairajkumar/covid19-in-india
[3] https://www.kaggle.com/sudalairajkumar/covid19-in-usa?select=us_states_covid19_daily.csv
[4] https://covidtracking.com/
[5] https://github.com/nytimes/covid-19-data

Challenges

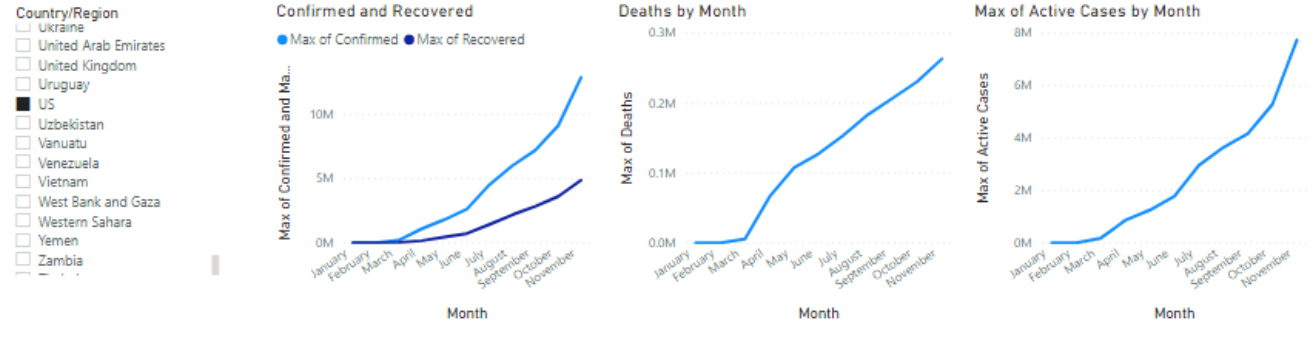
Some challenges that were faced:

- Preprocessing to infer important data. This included adding relevant state and coordinates to areas, presented weighted data to infer updated in Confirmed, recovered and deaths of individuals due to the virus.
- Presenting relevant data by identifying chart type and holding back on too many details. The dashboard presented only includes details needed from a coronavirus dashboard on a worldwide scale and on certain specific countries.
- Visual analysis was also a major challenge. This was resolved by looking at several BI guides on their website which have best practices.

Experiments

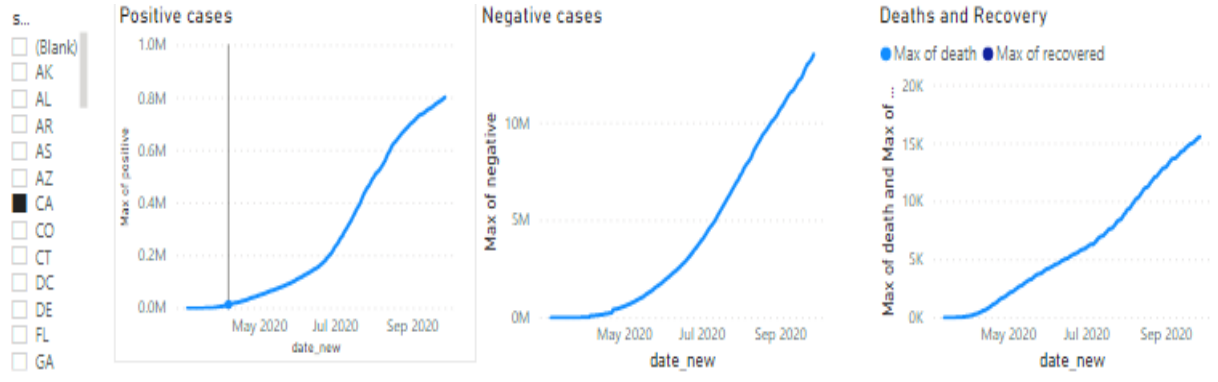
The theme of the project was to include a global perspective, followed by specific charts for some countries.

The very first chart helps us in selecting specific countries and finding trends in that particular country. This is followed by a geo map which helps in understanding coronavirus cases across the globe.



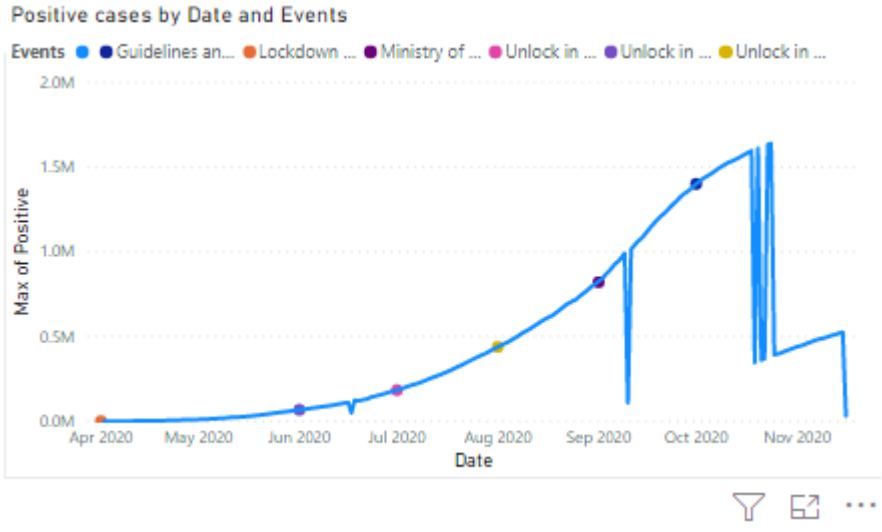
A case scenario is presented with weighted cases over 7 days, to get a better perspective as compared to daily changes.

In a similar fashion, US cases are presented where the reader can select States and have access to data of different states over the time period of the pandemic.



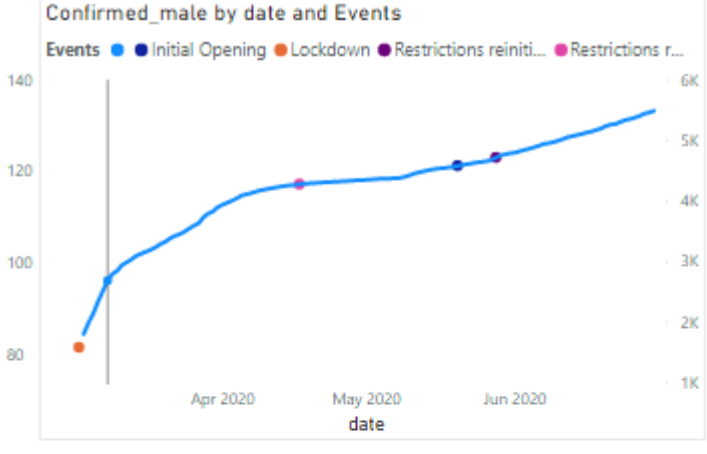
The US geo map presented gives us the trend of the virus as the reader can change the date and find the trend over time.

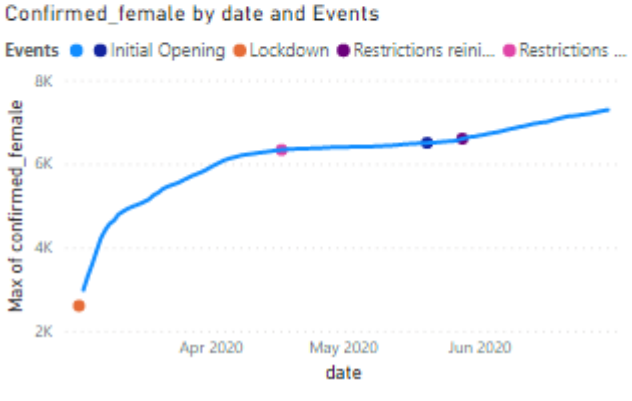
For India, the trend in coronavirus cases is presented with the legend for the different lockdowns that the country followed. Different countries faced the virus in different ways and these specific events help to understand the behavior of covid-19 cases as related to those events inherent to the country.



Similar to India, the analysis of South Korea is presented where the country had to reinitiate restrictions as cases rose when the initial lockdown was relieved.

Certain countries have more detailed data, with gender and age of patients presented as well. Additionally, the source of the virus is also classified such as foreign sources, cases pertaining to certain bars or restaurants, etc. These additional data help us in evaluating minor details about the virus and its spread.





Results

This project presented an exploratory visual analysis of the coronavirus pandemic across the globe. Several dashboards are presented which update in real time and the insight that such analysis presents help us in accessing data in the correct way.