COVID'19 DATA ANALYSIS

Shantanu Tyagi (201801015)* and Arkaprabha Banerjee (201801408)†
Dhirubhai Ambani Institute of Information & Communication Technology, Gandhinagar, Gujarat 382007, India
CS-306, Data Analysis and Visualization

In this report we analyze the progression of COVID-19 pandemic over time by taking various case studies and try to understand the effect of socio-economic and governmental interventions upon it. We shall furthermore analyze the general awareness among people in retrospect to the above parameters and try to qualitatively correlate it to standard epidemic spread models .

I. INTRODUCTION

Ever since of the inception of a local cluster of coronavirus in Wuhan, China in December 2019, the disease has progressed to be a full blown pandemic throughout the world. Transmission usually occurs via close contact or contact with the bodily fluids of an infected person. Different countries and regions have employed various strategies to contain the damage and spread inflicted by this pandemic. Effective analysis regarding these factors could potentially help prevent any such future circumstances .The behaviour of the general public also plays a pivotal role in this phenomenon as it helps us understand the effect awareness among them in regards to this pandemic.

In a bid to model epidemics, there have been multiple compartment models based on differential equations which seek to accurately define the trend based on the parameters. In this report we shall explore the SEIR model for SARS in relevance to our case studies. This SARS based model has been chosen on account of COVID-19 belonging to a similar class of disease.

All the relevant code can be found in the GitHub Repository

II. DATA ORGANIZATION

A. Dataset

In order to properly analyze the data the following dataset's were used:

- 1. COVID 19 Case data: Coronavirus Source Data - Our World in Data
- Government Measures: Blavatnik School of Government
- Search behavior — Google Trends

B. Data Features Used :

- New cases per day : New confirmed cases of COVID-19 (7-day smoothed)
- New cases per million population per day : New confirmed cases of COVID-19 per million population
- Total cases : Total confirmed cases of COVID-19
- Total cases per million population : Total confirmed cases of COVID-19 per million population
- New deaths per day : New confirmed deaths attributed to COVID-19 (7-day smoothed)
- New deaths per million population per day : New confirmed deaths attributed to COVID-19 (7-day smoothed) per million population
- Total deaths : Total confirmed deaths of COVID-19
- Total deaths per million population: Total deaths attributed to COVID-19 per million population
- Total vaccination per million population : Total number of COVID-19 vaccination doses administered per million people in the total population
- Reproduction Number : Number of secondary infections generated by an infected person in the time span for which they are infected.
- Stringency Index : Government Response Stringency Index: composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, re-scaled to a value from 0 to 100 (100 = strictest response)
- Governmental interventions : Consists of 4 major parameters each having multiple sub-parameters
  1. containment and closure policies
  2. economic policies
  3. health system policies
  4. miscellaneous policies

*Electronic address: 201801015@daiict.ac.in
†Electronic address: 201801408@daiict.ac.in

Each of their sub-parameters is rated from 0 to 3 where 3 represents urgent requirement in strict implementation of the policy and 0 represents a suggestion which may be implemented in the future.

For analyzing specific governmental policies country-wise the following containment parameters were used

1. School closing
2. Workplace closing
3. Cancel public events
4. Restrictions on gatherings
5. Restrictions on gatherings
6. Restrictions on internal movement
7. International travel controls

[3]

- Total Test's done : Total tests for COVID-19
- Search index Score : Numbers in the search index score represent the search interest relative to the highest point on the chart for the given region and time. A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular. A score of 0 means that there was not enough data for this term.

C. Handling Missing and irregular Data

- Missing data was encountered for the number of cases, number of deaths, number of vaccines and the reproduction Number. All other features had consistent data. Upon careful analysis it was found out that the missing data was present in the initial few days when there were no cases or deaths. Hence those fields were replaced by 0. Similarly for vaccinations every country started at a different point of time. Hence the initial few days have been allocated 0 whenever applicable.
- Missing data of the reproduction rate was filled by taking the value of the closest valid neighbouring field. This was done because reproduction rate is more or less close to it's neighbouring values on account of it's time dependent nature.
- For daily deaths, vaccinations and cases there could be often be systemic problems in reporting leading to high fluctuations, Thus a 7 day rolling average value has been chosen to handle such irregularities.

D. Case Studies

The aforementioned features have been studied with regards to 5 major countries and some qualitative discussions with regards to the data for the entire world. The 5 countries are :

- India
- Israel
- United States of America
- Italy
- New Zealand

For analyzing vaccination trends, data from 19-12-2020 to 03-05-2021 has been considered for all regions. For analyzing government intervention indices, data from 01-01-2020 to 04-05-2021 has been considered. For all other measures, data from 30-01-2020 to 21-04-2021 has been considered. The time series has been plotted in terms of number of days/weeks from the first relevant data point. While analyzing data across the above 3 groups with each other, the common subset has been considered.

III. SEIR MODEL

In the SEIR model, we have 4 major compartments : Susceptible, Exposed, Infected and Recovered. Furthermore, in order to model SARS more accurately, these 4 compartment have more sub-compartments as well. In this model we consider deaths as well unlike the SIR model. It furthermore assumes no births during the study period and the quarantine period as completely effective. The description of the subcompartments are given below :

- Susceptible (S)
- Susceptible Quarantined S_q : Susceptible's who are under quarantine
- Exposed (E) :Those who have SARS, but cant spread the infection yet.
- Exposed Quarantined E_q :Those who have the Disease, and are under quarantine.
- Infectious Undetected I_u : Have Infectious Disease but hasn't been detected.
- Infectious Quarantined I_q : Have Infectious Disease and are under quarantine.
- Infectious Isolated I_d : Have Infectious Disease and are completely isolated.
- Death (D): Death due to Disease.
- Recovered (R): Have Recovered from the disease and are now immune.

E , E_q , I_u , I_q , I_d are together considered as the net infected population.

$$\frac{dS}{dt} = u \cdot S_q - \frac{k \cdot I_u \cdot S \cdot (q + b \cdot (1-q))}{N}$$

$$\frac{dS_q}{dt} = \frac{k \cdot q \cdot (1-b) \cdot I_u \cdot S}{N} - u \cdot S_q$$

$$\frac{dE}{dt} = \frac{k \cdot b \cdot (1-q) \cdot I_u \cdot S}{N} - p \cdot E$$

$$\frac{dE_q}{dt} = \frac{k \cdot q \cdot b \cdot I_u \cdot S}{N} - p \cdot E_q$$

$$\frac{dI_u}{dt} = p \cdot E_q - m \cdot I_u - v \cdot I_u - w \cdot I_u$$

$$\frac{dI_d}{dt} = w \cdot I_u + w \cdot I_q - m \cdot I_d - v \cdot I_d$$

$$\frac{dI_q}{dt} = p \cdot I_q - m \cdot I_q - v \cdot I_q - w \cdot I_q$$

$$\frac{dD}{dt} = m \cdot I_u + m \cdot I_q + m \cdot I_d$$

$$\frac{dR}{dt} = v \cdot I_u + v \cdot I_q + v \cdot I_d$$

[2]
Various constants used in the above equations are defined as,

- b — Probability that a contact between person in I_u and someone in S results in transmission.
- k — Mean number of contacts per day someone from I_u has with someone in S.
- m — Per capita death rate.
- N — Initial population.
- p — Fraction per day of exposed people who become infectious. $\frac{1}{p}$ is the number of days in the early stages of SARS for a person to be infected but not infectious.
- q — Fraction per day of individuals in S that go into quarantine(S_q or E_q).
- u — Fraction per day of those in S_q who are allowed to leave quarantine back to being in S.
- v — Per capita recovery rate from different infectious categories to recovered.
- w — Fraction per day of those in I_u who are detected and isolated to I_d[2]

We have reproduction number(R) defined as,

- Reproduction Number (R) with no Intervention is :

$$R = \frac{k \cdot b \cdot (1-q)}{v + m + w}$$

- Reproduction Number (R) with Intervention is:

$$R = k \cdot b \cdot (1-q) \cdot D_{int}$$

where D_{int} is the mean duration of infectiousness under interventions

'q' represents the fraction of the susceptible population who shall undergo quarantine. Thus on increasing the value of q, we naturally find the population of the susceptible quarantined higher. This happens because the fraction of people who are quarantined don't contribute to the infected population as a result of which the number of deaths decreases as well. Peak infections increases as quarantine is delayed.

We observe that with a decrease in reproduction number the infected population also decreases. Thus in order to control the pandemic we need to reduce the reproduction number. Bringing it to a value less than 1 helps us curtail the effect of the pandemic. On account of various social and governmental interventions, the final reproduction number (R_{int}) reduces. When we are able to bring it below 1, that's when the epidemic is successfully tackled, as the infectiousness of the disease reduces. The epidemic ends when the number of infected becomes 0.

'k' is the mean number of contacts someone from the infectious undetected compartment has with the susceptible population. When the number of contact is significantly low, we hardly observe any deaths or infected people. But on increasing the value of k, we find that the number of deaths and the infected population has increases rather significantly. This is because with the increase in number of contacts per person, the disease spreads faster.

There exists no closed form expression of the above model. Hence we try to qualitatively analyze the trend for real life scenarios and compare those to representative models.
For the purpose of this report we shall only look at the number of infected and the number of deaths. Number of infected corresponds to the number of new cases per day whereas the the number of deaths correspond to the total number of deaths up till that point.

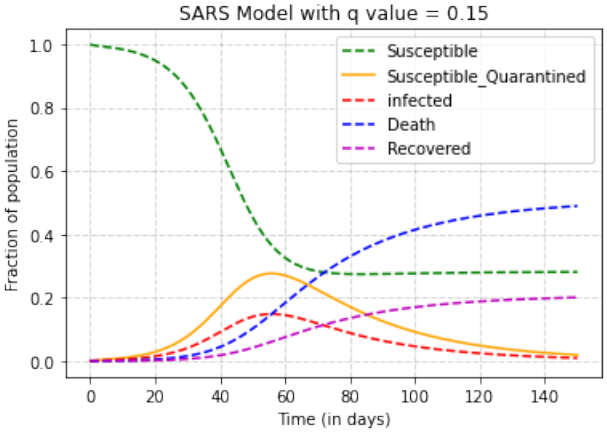


FIG. 1: Representative plot for the model

IV. OBSERVATIONS

A. Overview

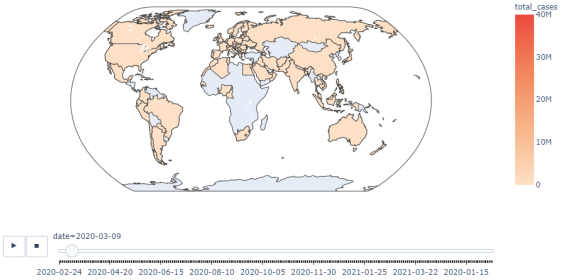


FIG. 2: On 09-03-2020

The figure shown above represents the total cases registered in each country during the onset of the pandemic.

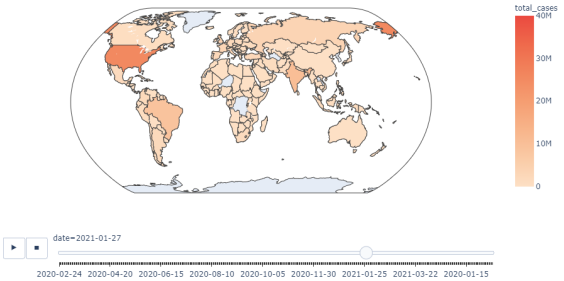


FIG. 3: On 27-01-2021

The figure shown above represents the total cases registered in each country up till a much recent date.

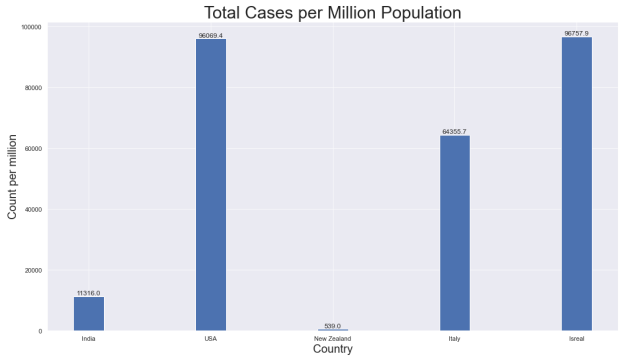


FIG. 4: Cases per million population

The above plot shows the cases per million population in the 5 countries that we have considered for our study.We shall now analyze the Correlation between total number of tests, total number of positive cases and total number of deaths for the entire world.

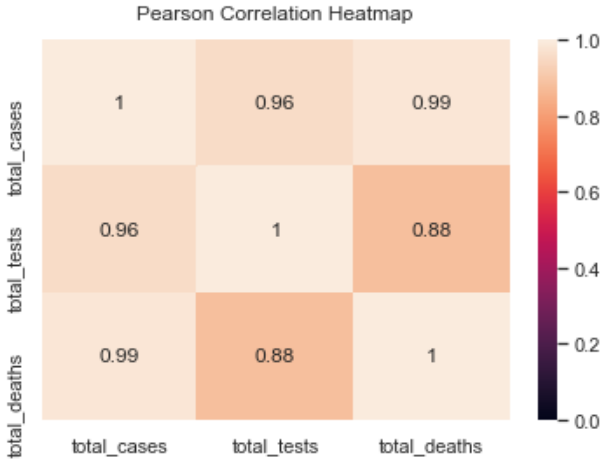


FIG. 5: Pearson correlation matrix for cases, tests and deaths

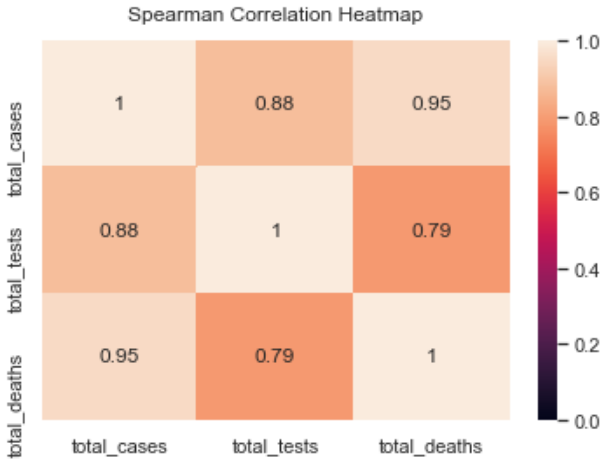


FIG. 6: Spearman correlation matrix for cases, tests and deaths

The above 2 correlation matrices for the entire global data show a strong correlation between the number of tests done and the number of people who were found positive. This shows the need for more and efficient testing so that the countries can get reliable infection estimates and can be better equipped in coordinating containment and relief efforts.

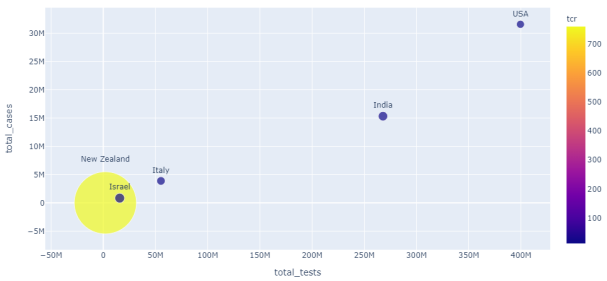


FIG. 7: Cases Vs Tests scatter plot

The above graph shows a scatter plot between the total cases and total tests for 5 countries.

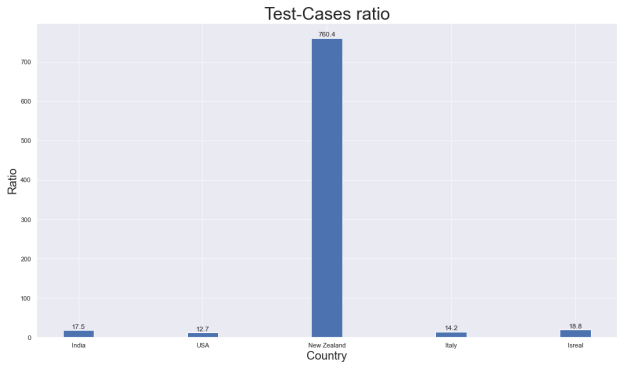


FIG. 8: Test to Cases Ratio

The above graph shows a plot of total tests to total cases for 5 countries.
New Zealand has been attributed to be one of the best countries to handle this pandemic. The primary reason can be seen above. It has a significantly higher test to confirmed ratio ≈ 760. We can observe most of the other countries have a significantly lower ratio of around 18. This means in order to properly get an idea and track the spread of cases in the population the government should ideally perform 760 tests for every single infected person.

B. Number of cases

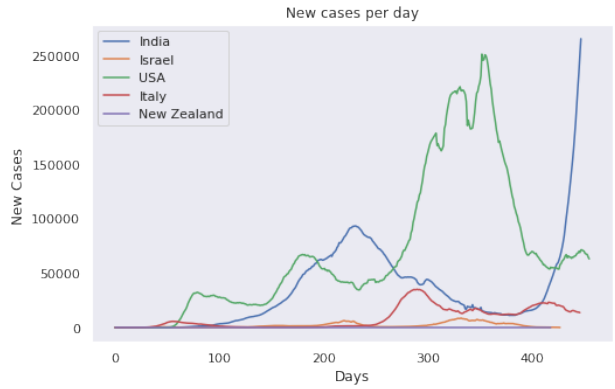


FIG. 9: New Cases

For new cases per day we should ideally get a single peak for the number of infected people as the number of infect should tend to 0. But here we observe multiple peaks. This is because as soon as the number of cases per day goes down people become extremely careless (can be seen from stringency index) and thus the number of infected starts to rise again.

We also observe the following :

1. India had previously flattened the curve but is now rising at a huge pace. India also has a single peak till date.

2. Italy has a late peak because initially the government had put in stringent measures but later on people started disregarding it hence the increase

3. USA has multiple peaks because everytime the infections started to go down, restrictions were lifted and there was utter disregard for protocols

4. Israel and New Zealand have comparatively low populations hence the number of new cases is negligible

We shall now look at the Number of new cases per million population.

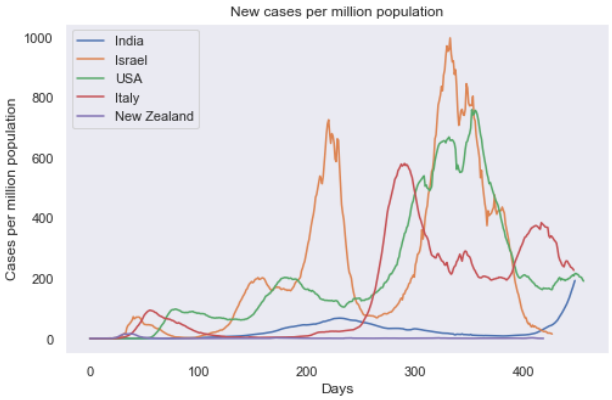


FIG. 10: New cases per million population

We observe the following :

1. We find Israel to be the fore-runner here. This is on account of the fact that Israel achieved herd immunity thus there were larger number of cases per million population.

2. Although India had a very low number of cases per million but lately it has been rising.It is especially worrisome due to the extremely high population density and the lack of proper healthcare for such a large population.

We shall now look at the statistics for total number of cases

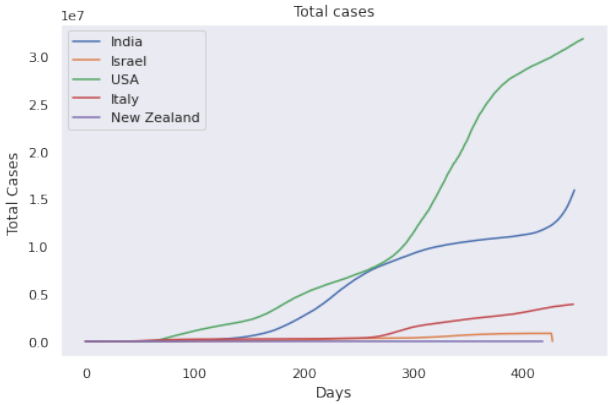


FIG. 11: Total cases

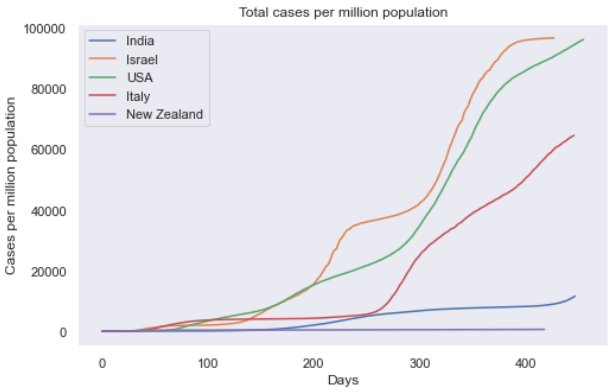


FIG. 12: Total cases per million population

One can observe the following :

1. Total cases in USA and India surpass others by a huge margin primarily due to high population
2. Israel again has a very high number of total cases per million population due to it achieving herd immunity.

C. Number of deaths

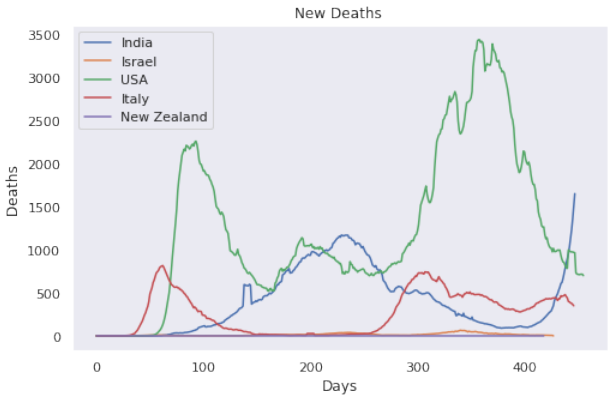


FIG. 13: New deaths

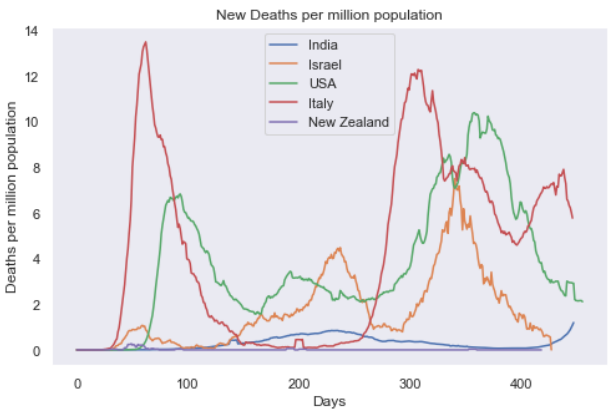


FIG. 14: New deaths per million population

We observe the following for New deaths :

1. USA has 3 peaks each corresponding to the rise in new cases with some time lag.
2. Italy has higher number of Deaths per million population. It also has 2 peaks. The first peak is on account of the fact that Italy was one of the first countries to be hit badly and they weren't equipped for it. The second peak in death corresponds to the relaxation of containment measures.
3. New Zealand has negligible number of deaths
4. India has low number of deaths per population but a significantly higher number of absolute deaths per day. This is on account of the high population density.

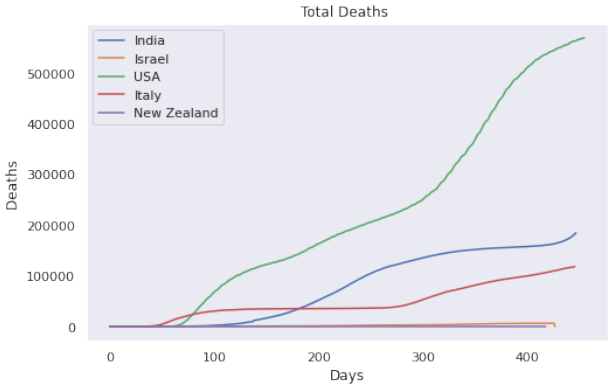


FIG. 15: Total deaths

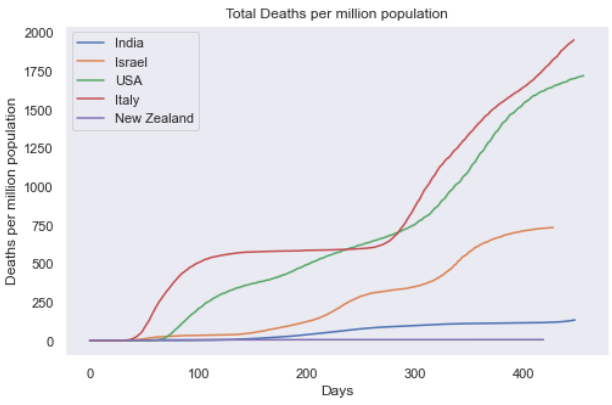
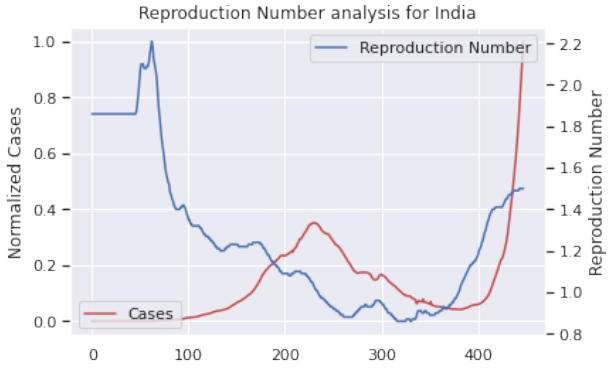


FIG. 16: Total deaths per million population

One can also observe that even though the number of cases per million in Israel were very high but the number of deaths per million is quite low thus confirming the fact that they are indeed achieving herd immunity. Furthermore Italy has an extremely high number of Deaths per million population. India and USA still have an extremely high number of total deaths due to its population.

D. Reproduction Number Analysis

The variation in Reproduction number has been performed for India. It is being compared to the normalized number of new cases per day (value between 0 to 1).



number for India

One can observe that as long as the reproduction number stays above 1 we witness an increase in new cases per day. However, as soon the number falls below 1 the number of cases start dropping.

E. Vaccination

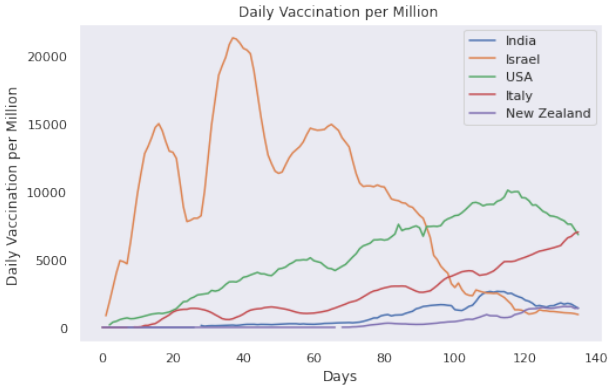


FIG. 17: Daily Vaccinations per million population

Israel was one of the first countries to start vaccination aggressively and had the highest number of vaccinated population. This couple with the herd immunity that they were achieving helped control the pandemic in this country. Other countries are slowly starting adopting vaccinations to control the pandemic. An interesting point to note is that New Zealand has an extremely low number of vaccinated population and yet it has the least number of cases and deaths per million.

F. Stringency Index Analysis

We shall now analyze the stringency index in relation to the Normalized number of cases (Scaled down to 100) country-wise.

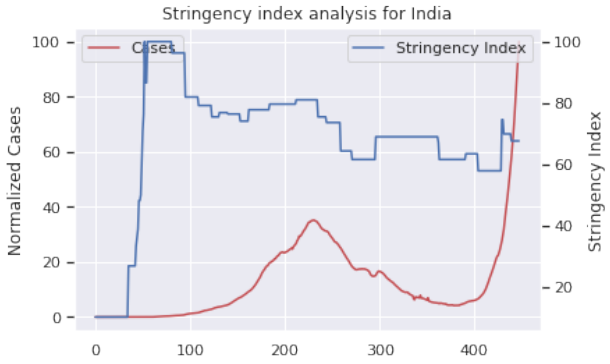


FIG. 18: Stringency Index for India

India initially had stringent measures in places which helped check the growth rate and we assumed the peak had been reached. But as the stringency index further started decreasing we experienced a huge rise in new cases per day due to lack of proper measures.

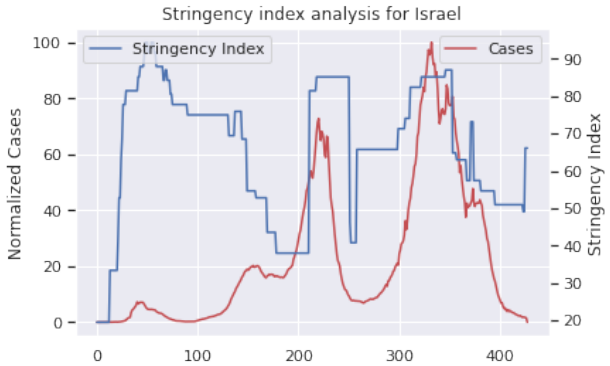


FIG. 19: Stringency Index for Israel

Israel followed an alternate strategy. It increased it's stringency index to a very high level the moment cases started increasing but relaxed it as soon as the cases got down.

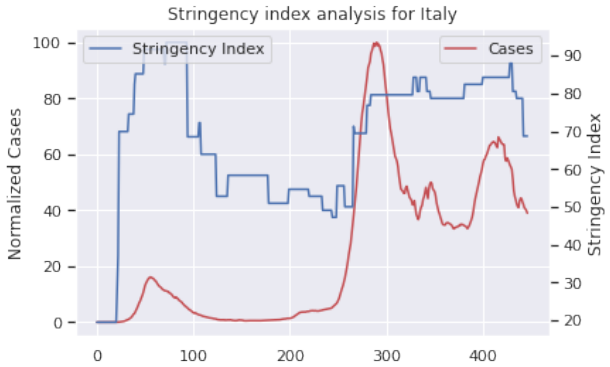


FIG. 20: Stringency Index for Italy

Italy was one of the first countries to be hit by the pandemic. Hence it elevated all possible measures to the highest level to control it. This turned out to be quite successful. However the moment the cases

started dropping, the measures were relaxed and cases started rising exponentially.

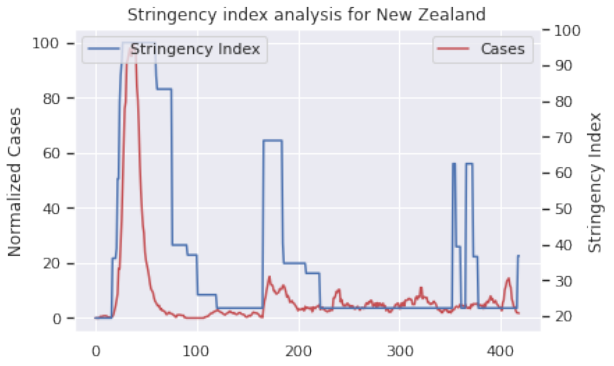


FIG. 21: Stringency Index for New Zealand

New Zealand perhaps had one of the best governmental approaches. Even before the onset of the pandemic in New Zealand they had the highest level of containment measures already implemented. This is in stark contrast to other countries which implemented Measures after they had been hit by this pandemic. Although countries like Italy and India also had higher stringency indices in the initial duration yet New Zealand had the most effective outcome. This is primarily because New Zealand had the highest level of containment measure for international trips. Since the pandemic was not local to the country thus by limiting the number of infected people entering into the country, they managed to control the spread. Most of the other countries implemented ban on International Flights after a while and focused more on internal containment measures.

G. Governmental Interventions

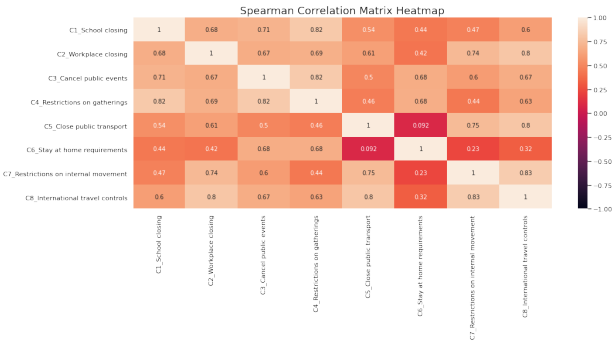


FIG. 22: Feature Correlation matrix

The above correlation matrix has been generated via the data of the aforementioned countries. Following are the major observations :

1. With the closing of international transport, work-spaces and public transport also closed down. Restrictions on internal movement also increased.
2. Restrictions on Public events and restrictions on gatherings are heavily correlated
3. Stay at home requirement increased along with shutting down of schools with increasing restrictions on public gatherings.

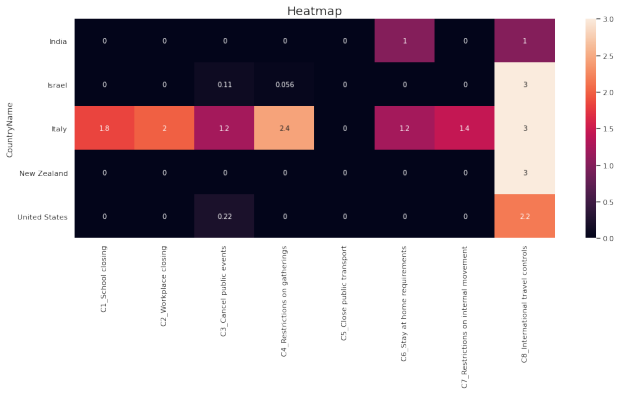


FIG. 23: Containment Measure Value

The above heatmap shows the average government response index from 15th February 2020 to 28th February 2020. By this time apart from New Zealand, other countries had started registering cases. One can observe that New Zealand already had Strict International travel protocols even before it registered a single case. Italy shows other containment measures as well because it was already registering a large number of cases.

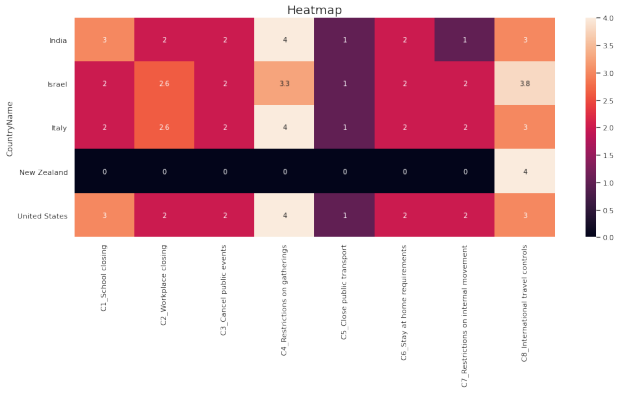


FIG. 24: Containment Measure Value

The above heatmap shows the average government response index from 15th December 2020 to 1st January 2021. We observe that New Zealand only has strict international travel protocols and no other containment measure even when most other countries were applying all possible measures.

1. The success of New Zealand in controlling the pandemic is due the government giving utmost priority in limiting international travel even before the pandemic started there. Thus once the limited number of cases which occurred were controlled the government relaxed all internal measures. Workplaces, Schools and other public institutions worked normally.

2. Most of the other countries fixated upon measures once the disease started spreading in the country. Thus rather than international travel measures, they were more fixated upon Internal measures. For countries like India having a very high population density, this led to work from home practises and shutting down of schools and public events and institutions. Since the disease had already spread enough thus even a slight relaxation led to severe increase in cases.

3. With proper containment measures and efficient lockdown we can reduce the average number of contacts per person which in turn leads to a decrease in reproduction number.

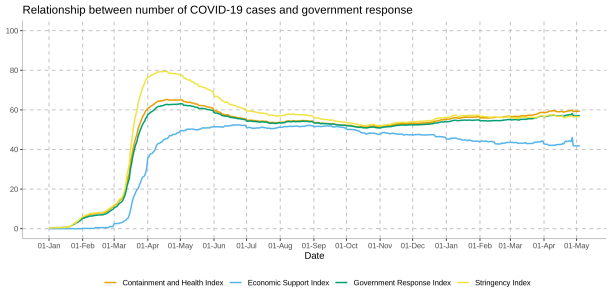


FIG. 25: Overall index value

We can further analyze that the economic support provided by governments all over the world is quite less than the measure it attributes to handling containment or health measures. Thus one may notice that poor or underprivileged people may be severely affected by these measures.

H. Search Trends Analysis

Here, we try to analyze the pattern between google searches related to COVID'19 and its vaccine to the total cases on a weekly basis. The cases have been normalised to 100 since the tool returns normalised values. We assume that google search trends directly translate to the net awareness among people.

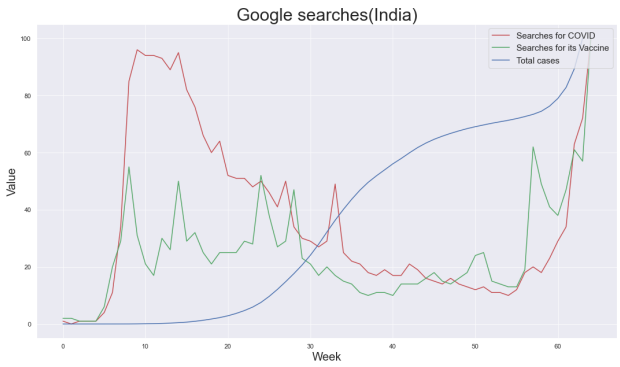


FIG. 26: Google searches in India

In case of India, we observe that during the onset of the pandemic, the searches involving the coronavirus were very high while the cases were very less. The search results gradually died out as the COVID cases in India started saturating. At this point people began looking up for vaccines. However India entered a second wave with a much faster spread resulting in both virus and vaccine related searches going up very quickly.

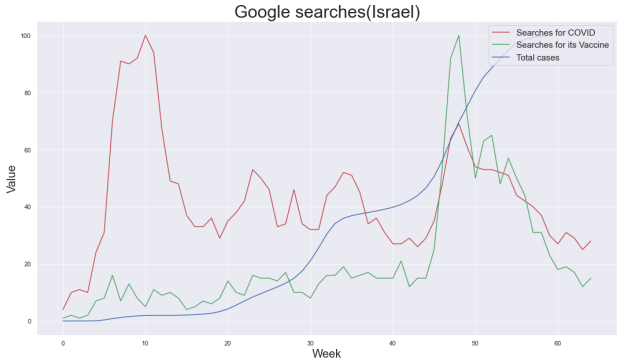


FIG. 27: Google searches in Israel

In case of Israel, we observe a similar trend for the coronavirus search results during the onset of the pandemic however there were not many searches for vaccines. As the cases increased, we saw a peak in the virus and its vaccine related searches which later reduced as situations improved.

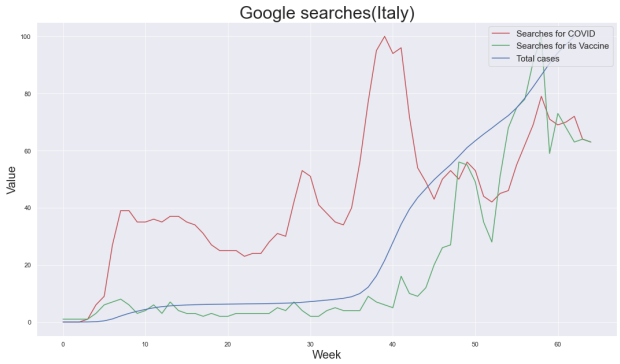


FIG. 28: Google searches in Italy

In case of Italy, there were very few searches for both these categories which have increasing with local peaks with weeks when the total cases suddenly increased.

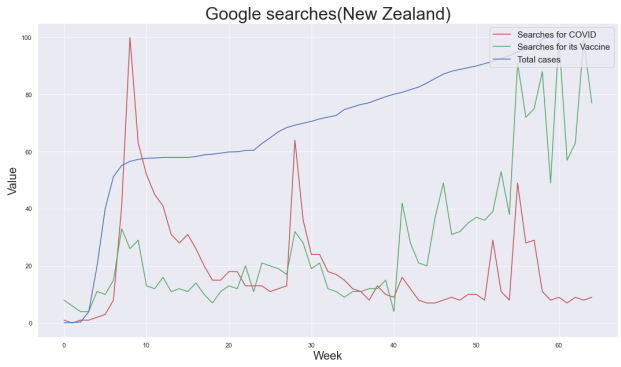


FIG. 29: Google searches in New Zealand

In case of New Zealand, the initial searches were very high for the virus related information and those regarding the vaccines were higher than other countries. This was because of aggressive awareness campaigns and containment measures by the government. Coronavirus searches have been decreasing with time as the total cases saturate while that of vaccines have been increasing like that in case of India.

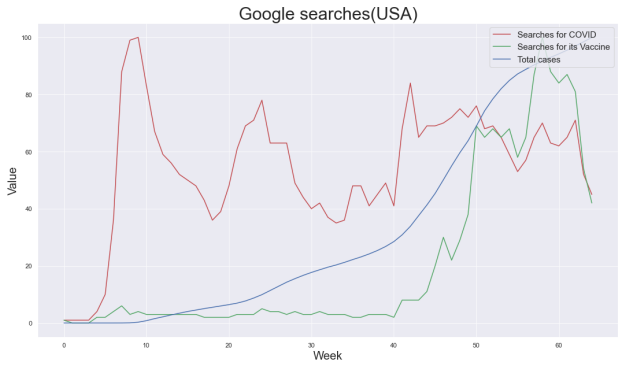


FIG. 30: Google searches in USA

In case of USA, we observe that the coronavirus related searches have stayed fairly higher and much consistent as compared to other countries from the starting itself. The vaccine related searches have been increasing with the number of cases.

In most of the above cases we observe an increase in searches as the rate of increase of total number of cases increased or when the government announced containment measures. Thus the way a particular government handles a pandemic directly affects the awareness and search patterns of its citizens.

V. CONCLUSIONS

We analyzed and visualized the the data from the COVID'19 pandemic and related it with the SEIR compartment model. In this model we encounter only single peak since the infections die out however in real life we can see multiple peaks which can be locally explained by the irresponsible human behaviour. We did time based visualisation for the cases all around the world and plotted the correlation heat-maps between cases, tests and deaths to find monotonic relationships between these features for accumulative data of all countries . We selected 5 countries, namely, USA, India, Italy, Israel and New Zealand by considering how these countries handled the situations using a different approach than the other countries. We made several visualisations not only on the objective data but also on subjective data like stringency and interventions made by the government in a bod to explain spread of the virus in the countries of our interest. We found out that the ideal measure to control a pandemic which hasn't originated in that country is to control the inflow of the disease from international sources before it can spread. For regions of high population density, continued and effective containment protocols must be ensured else we may find ourselves going through another recurrence of the past episodes. On the other hand, for regions of low population density, containment protocols may be modified as per requirement. Moreover, Vaccination drive must be started as soon as possible to control the reproduction number. Maximum number of tests should be done as soon as possible to garner an idea about the extent of the pandemic so that we can take the required steps. We saw that in India, even though we have low number of deaths per million and cases per million, but the healthcare system doesn't scale up in the same manner hence we find ourselves in a problematic scenario. We furthermore, tried to analyze how awareness among the general population varies as the pandemic progressed by collecting the data for google search results. We observed that the search popularity increased with an increase in the number of cases or with governmental initiatives. However the exact pattern depends upon the region of the interest.

[1] Lipsitch M, Cohen T, Cooper B, Robins JM, Ma S, James L, Gopalakrishna G, Chew SK, Tan CC, Samore MH, Fisman D, Murray M. Transmission dynamics and control of severe acute respiratory syndrome. Science. 2003 Jun 20;300(5627):1966-70. doi: 10.1126/science.1086616. Epub 2003 May 23. PMID: 12766207; PMCID: PMC2760158.

[2] Module 6.2, A. Shiflet and G. Shiflet, *Introduction to Computational Science: Modeling an Simulation for the Sciences*, Princeton University Press.3, 276 (2006).

[3] Hale, T., Angrist, N., Goldszmidt, R. et al. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). Nat Hum Behav 5, 529–538 (2021). https://doi.org/10.1038/s41562-021-01079-8