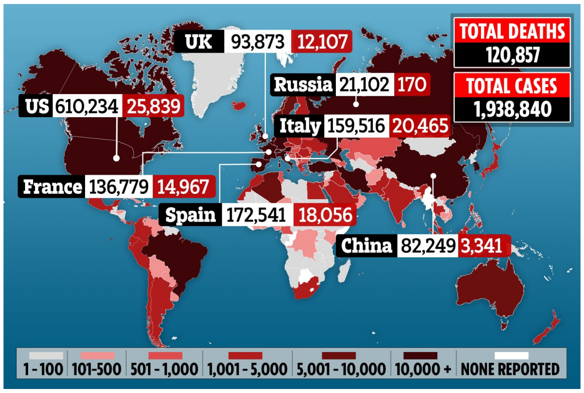
Covid-19 Spread Analysis Worldwide

Abhilash Kumar, *M.Sc. Big Data Analytics and AI*, LYIT, Letterkenny, Ireland

***Abstract*—Since its outbreak in Dec 2019, Covid-19 Coronavirus has spread in all countries worldwide in just a span of few months time. It has drastically impacted all aspects of life. With first case reported in Wuhan city of China in December 2019, it has quickly spread all over the world. As of March 2021, the confirmed cases and deaths are still getting reported on a daily basis. Although vaccines are developed to encounter the situation, still there are much to do to control the situation. As of 3rd March 2021, there has been 115M reported cases and 2.54M deaths due to the deadly virus. As the world's complete attention today has been towards the spread and impacts of Covid-19, this work has been carried out to analyse the covid-19 data on the number of infected as well deaths worldwide and to make it understandable using Artificial Intelligence and Big Data Analysis. We will also try to build AI model using Machine Learning techniques which should be able to predict the number infected people based on the trend available.**

With great advancements in technology, Artificial Intelligence and Big Data Analysis are ready to take any challenges posed to the Business and human life due to the Pandemic situation. Our expectation from this work is to provide better insights from the existing data and improves Covid-19 situation.



I. Introduction

CORONAVIRUS has changed the entire life style of the world drastically. Every individual has been impacted by this novel coronavirus. With a major impact of health care domain, it has equally impacted other areas of human life such as businesses, education, transportation, constructions etc. Due to its human to human transmission, Covid-19 has become much more dangerous compared to other members of coronavirus family. It has spread very rapidly in all the countries in the world. As of first week of March, 2021, United States of America has the most number of infected cases which stands around 29M. Some other countries like Brazil, India, Italy are also enormously impacted. With few vaccines in place, it is still spreading on a daily basis.

As a preventive measure, WHO and national governments have issued public advices and guidelines to fight against the spread of the virus. Many countries has to implement complete lock down for months to control the spread of deadly virus. In the effort to prevent the infection, normal life has been stalled, which has impacted people in many different ways.

For the purpose of above stated challenge, we are planning to analyse the spread of the virus in different countries since it has been reported. We are going to use 3 datasets of confirmed cases, deaths and recovered for all the countries since 22-Jan-2020. As part of this initiative, we are going to use the power of AI and Big data tools to

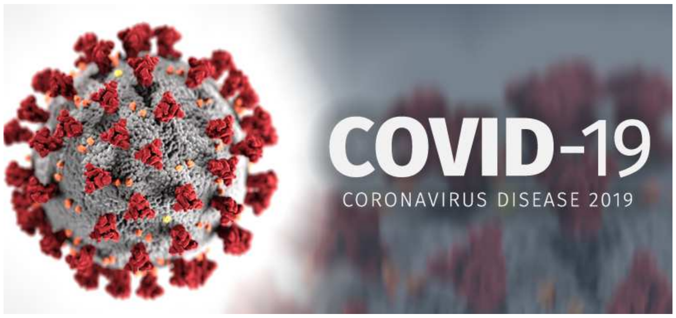


Fig. 1. COVID-19.

Very challenging thing in the data analysis is to collect the correct data. There are altogether 3 different phases while collecting the data. Starting with doing everything manually, we can perform data collection in semi automated way and fully automated way. Normally fully automated way of data collection is performed when the project is in much matured stage. Manual gathering of data involves methods like web scraping, where as CSV retrieving is semi automated and third phase is API calls.

Here, in this work we are going to use time series data of confirmed cases, deaths and recovered cases available in Johns Hopkins GITHUB. We are going to first read the three CSV files into our Spark data frame. After we reorganise the data in a format so that we can analyse it, we are going to visualize the data so that it is much more understandable. Next we will plot the infected Vs deaths graphs to visualise the relative correlation between them. Finally, we will prepare a predictive model to predict any trend in a particular country.

Rest of the paper below is organized as follows: Section II illustrates some basic knowledge of Covid-19 and motivation

Under the supervision of Dr. Shagufta Henna, Letterkenny Institute of Technology, Letterkenny, CO. Donegal behind the use of Big Data and AI. Section III, Dataset which describes our dataset as a whole. All the features available and then any challenges before we can start its processing. Section IV would describe our main approach toward data analysis. This section will typically includes Data Acquisition process, Data Cleaning process, Data Enrichment process and then finally Plotting/Visualization and Predictions. Under prediction, we will use linear regression models to predict the trends. In the next Section V in this paper, we will evaluate our predictive model. In the final section VI of this paper, we will discuss few ideas that can be implemented in future to get better analysis options. These ideas will be mainly governed by the experience achieved while performing the current task.

II. Covid-19

Coronavirus (Covid-19) is a virus that can affect lungs and respiratory system. The first case was reported in Wuhan city of China in Dec 2019. Since then is has rapidly spread in all countries in the world. Despite of few vaccines in the market, public guidelines and partial lock downs world wide, there are still very large number of new infected cases coming up on a daily basis. Deaths toll is almost 3% of the infected cases.

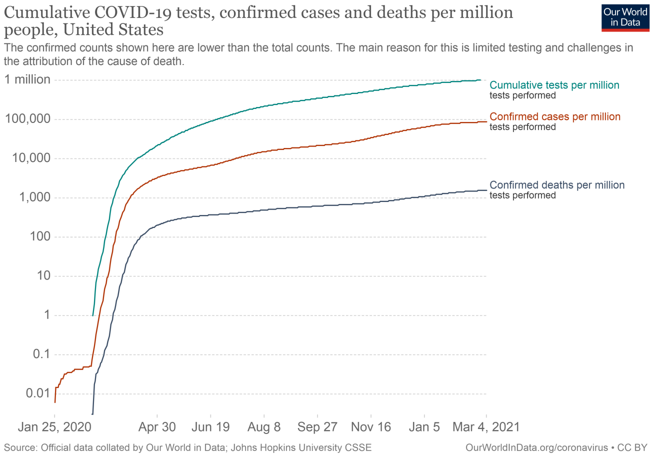


Fig. 2. Tests-Cases-Deaths(Based on Log)

The pandemic has impacted the world in such a way that all eyes are focused on to prevent its spread. Many attempts has been made to fight against the outbreak. National governments efforts are mainly targeted to stop the pandemic with different measures, making sure to have their hospitals and healthcare staffs ready for any crisis situation and provide supports to their economy by providing emergency packages. On the other hands Individuals responsibility is it adhere to all the public rules during this situation for example wearing face masks at public places, frequently washing hands, maintain social distancing and avoid any unnecessary travels. In line with these efforts, Research and Development has also prioritized relevant to Covid-19.

Computer science has also made a tremendous efforts during this period to help fight against coronavirus. With AI and Big Data making impacts in almost all areas today, it has helped a lot in understanding the virus behaviour and its spread. With this project, we are going to understand the spread of the virus since it has started in all the countries and going to predict the numbers based on the data available. Our project is supported by lot of data available today related the numbers pertaining to confirmed cases, recovered cases and number of deaths county wise.

Here we are going to use Spark in Databricks to capture and process our data. We are using Python as a programming tool to achieve our objective.

III. Dataset

To continue the work on such projects involving predictions, we need sample data where we can perform our analysis using different Big Data Analysis tools and Machine Learning techniques.

Data collection is the back bone of any Big Data projects. Right amount of data plays a significant role in the success of any such projects. We have used Johns Hopkins data source to gather the time series data of Covid-19 cases. Johns Hopkins [3] is collecting data in a daily fashion and everyday they are updating the data. It is located in its github location. We have total 385 columns and 193 rows in all the 3 datasets which contains numbers from 183 different countries. Below are three CSV files that we are going to use in our analysis.

- **time_series_covid19_confirmed_global**
- **time_series_covid19_deaths_global**
- **time_series_covid19_recovered_global**

We have below columns in our original dataset:

- Province/State: This column represents basically the area within a particular country.
- Country/Region: This is the actual county name.
- Lat: Latitude of the county location.
- Longitude: Longitude of the country location.
- Dates: This column is basically all the dates ragning from 22nd jan 2020 until the most recent one when the data is retrieved from the Johsn Hopkins github location. Under each date we have a number indicating the confirmed case or deaths or recovered case based on the file we are looking into

Each columns in the data sets are inspected for if they are useful in the analysis. There are some columns that are not required in our analysis, those columns are dropped up front so that data is more cleaner for the next level of processing. We have used python code to remove the columns and data cleaning.

IV. Main Approach

Application of Big Data in fighting with Covid-19 has been proven phenomenal so far. Some key concepts of Big Data Analytic such as multi-domain dataset analysis, deep analysis, handling of high dimensional data and parallel computing has been quite beneficial in terms of outbreak predictions, spread analysis, diagnosis and treatment.

In this section of Covid-19 data analysis, we are going to process the data in various steps. Fig.3 below details the different processing stages of the data. [1]

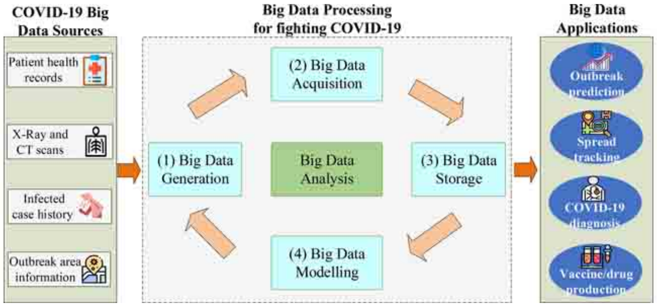


Fig. 3. Covid-19 Data Processing.

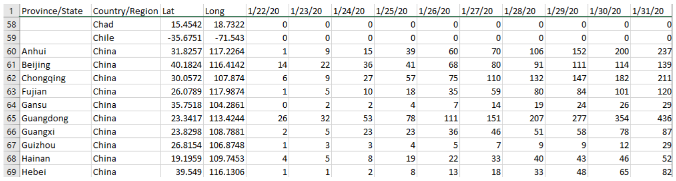
Data Acquisition:

This step of data preparation is vital as we need to identify right data for the analysis. As discussed above, we have identified Johns Hopkins covid time series data for our analysis. The data is stored in 3 CSV files containing confirmed cases, deaths and recovered cases separately. Our goal will be to analyse these data country wise and visualize any trends in the data. We have directly copied the files and stored in Databricks file repository.

Data Cleaning:

This is very important step in data Analytic projects. As part of this step, we need to explore data keeping in mind the original goal and identifying how different features are linked together to get achieve the goal. Some features might be of no importance to our analysis. At the same time others might have wrong entries or even they are missing. We need to make sure these aspects are adhered so that our data become homogeneous and clean.

The data retrieved in our case has below format.



1	Province/State	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	1/31/20
58		Chad	15.4542	18.7322	0	0	0	0	0	0	0	0	0	0
59		Chile	-35.6751	-71.543	0	0	0	0	0	0	0	0	0	0
60	Anhui	China	31.8257	117.2264	1	9	15	39	60	70	106	152	200	237
61	Beijing	China	40.1824	116.4142	14	22	36	41	68	80	91	111	114	139
62	Chongqing	China	30.0572	107.874	6	9	27	57	75	110	132	147	182	211
63	Fujian	China	26.0789	117.9874	1	5	10	18	35	59	80	84	101	120
64	Gansu	China	35.7518	104.2861	0	2	2	4	7	14	19	24	26	29
65	Guangdong	China	23.3417	113.4244	26	32	53	78	111	151	207	277	354	436
66	Guangxi	China	23.8298	108.7881	2	5	23	23	36	46	51	58	78	87
67	Guizhou	China	26.8154	106.8748	1	3	3	4	5	7	9	9	12	29
68	Hainan	China	19.1959	109.7453	4	5	8	19	22	33	40	43	46	52
69	Hebei	China	39.549	116.1306	1	1	2	8	13	18	33	48	65	82

Fig. 4. Covid-19 Data Structure.

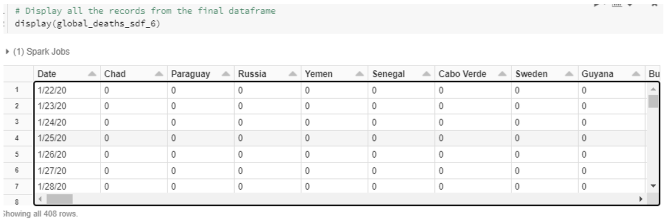
The values under column 'Province/State' has some missing values. Ideally we could have dropped those values however that could result in major data loss. At the same time we don't really need few columns in our analysis. So we have decided to drop them from our dataset. For example, we are dropping the columns **'Province/State', 'Lat',** and **'Long'**.

Data Enrichment:

In the data enrichment process, we are manipulating data in such a way to get most value out of it. We need to reformat and narrow down the data to an extent that we are left with only important data which should contribute to our analysis.

In our present case study we have dates in the header. The present information in the header are of different types. Its really tough to analyse data in this format. Also when we read this data in a dataframe and try to analyse it, we should have the dates in date type. So we have a challenge to reformat this into new data structure where we must have all dates under a date type column. In the new format, countries will become the header where as values under each country will represent corresponding number for a particular date.

Example of the data set ready for the visualization and processing is as below:



```
# Display all the records from the final dataframe
display(global_deaths_sdf_6)
```

(1) Spark Jobs

	Date	Chad	Paraguay	Russia	Yemen	Senegal	Cabo Verde	Sweden	Guyana	Bu
1	1/22/20	0	0	0	0	0	0	0	0	
2	1/23/20	0	0	0	0	0	0	0	0	
3	1/24/20	0	0	0	0	0	0	0	0	
4	1/25/20	0	0	0	0	0	0	0	0	
5	1/26/20	0	0	0	0	0	0	0	0	
6	1/27/20	0	0	0	0	0	0	0	0	
7	1/28/20	0	0	0	0	0	0	0	0	
8										

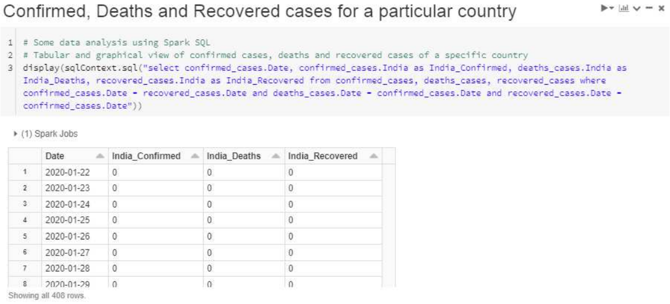
Showing all 408 rows.

Fig. 5. Final Dataset.

Plotting and Visualization:

Once we are done with data preparation, we need to start data exploration. Typically data exploration is done through visual manner. In this section we have used built in visual libraries to plot the data from the CSV file. Mostly Matplotlib is used to visualize the data. First of all we are going to visualize the trends of confirmed cases, deaths and recovered cases for a particular country. Here we have used Spark sql utility on our 3 datasets to bring the required fields in a table for all dates and then visualize the trends using the Databricks inbuilt utility. We have chosen to use the line graph as it clearly shows how the 3 lines are plotted with time. Also the death line is closer to X-axis, which says the number of deaths in India due to Covid-19. At the same time the confirmed cases and recovered cases are trending closer meaning that the numbers are closer to each other with time and the recovered rate is better.

The tabular form of the aggregated data is shown in the Fig.6 below



Confirmed, Deaths and Recovered cases for a particular country

```
# Some data analysis using Spark SQL
# Tabular and graphical view of confirmed cases, deaths and recovered cases of a specific country
display(sqlContext.sql("select confirmed_cases.Date, confirmed_cases.India as India_Confirmed, deaths_cases.India as
India_Deaths, recovered_cases.India as India_Recovered from confirmed_cases, deaths_cases, recovered_cases where
confirmed_cases.Date = recovered_cases.Date and deaths_cases.Date = confirmed_cases.Date and recovered_cases.Date =
confirmed_cases.Date"))
```

(1) Spark Jobs

	Date	India_Confirmed	India_Deaths	India_Recovered
1	2020-01-22	0	0	0
2	2020-01-23	0	0	0
3	2020-01-24	0	0	0
4	2020-01-25	0	0	0
5	2020-01-26	0	0	0
6	2020-01-27	0	0	0
7	2020-01-28	0	0	0
8	2020-01-29	0	0	0

Showing all 408 rows.

Fig. 6. India Confirmed Recovered deaths cases.

Fig.7 shows the visualization of the same table which depicts the trends of the three categories in the period starting from 22-Jan-2020. The graph is exponential graph and it could be better represented using log values of the number of cases. We will use logarithmic values further down the line to represent the trends.

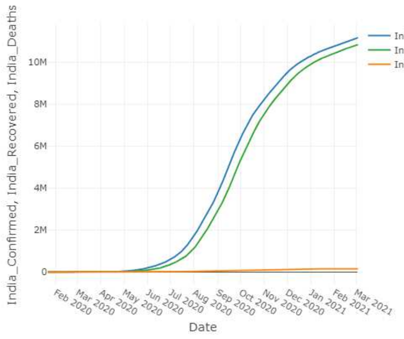


Fig. 7. India Confirmed Recovered deaths cases.

Our next step in the visualization is to see the trend of confirmed cases in few selected countries. For this purpose we have taken 5 countries into consideration. And used Plotly object to plot the graph. Below Fig.8, Fig.9 and Fig.10 shows the trends of 3 different categories which is actually 3 different datasets.

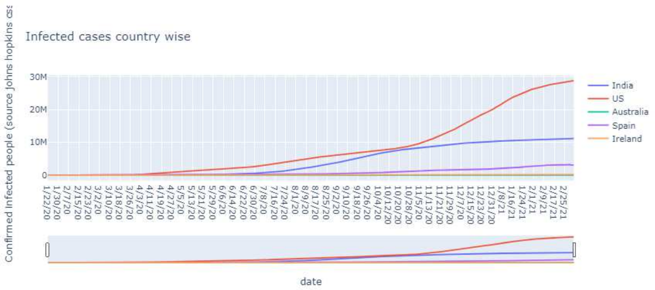


Fig. 8. Infection Trends Worldwide.

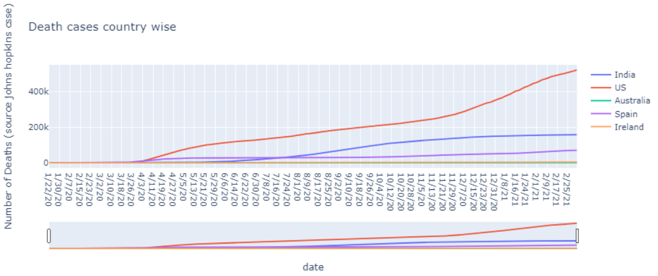


Fig. 9. Deaths Trends Worldwide.

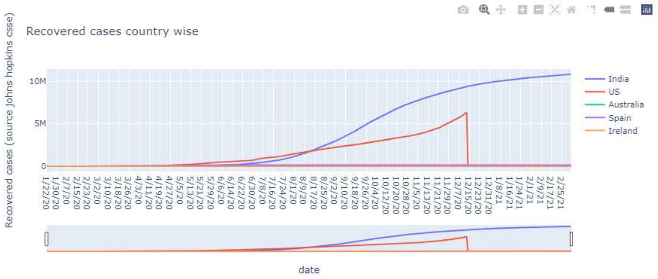


Fig. 10. Recovered Trends Worldwide.

Our next analysis is to find the doubling rate determination of the trends for a specific country. We would like to know which country has doubling infected rate in 2 days or 5 days or 10 days. Doubling rate 2 means every 2 days number of infected people doubles. Same way we can determine the doubling rate for deaths and recovered cases. We need here some helper line to determine the current slope of each trending line. For this purpose we first determined the doubling rate on a base value of 100 using active days of 15 and plotted a graph along the trending graphs. This way we can easily identify the trends having their doubling rate.

Below equation is used to calculate the doubling rate on a base of 100 and included in the dataframe so that it can be plotted along with the different trends for easy identification.

$$N(t) = N_0 * 2^{t/T} \tag{1}$$

In the Fig.11, it clearly shows that the Infected rate in US is quite closer to doubling rate of 2 days. At the same time India and Australia are closer to the doubling rate of 5 days in terms of infected people.

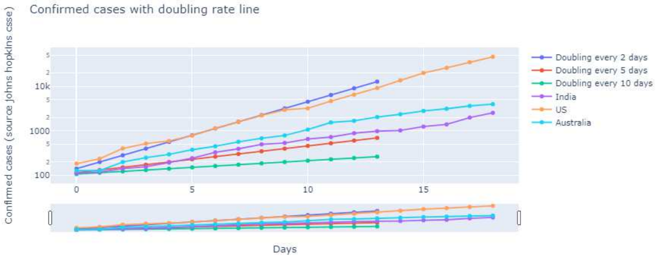


Fig. 11. Confirmd Cases Doubling Rate.

In the same way below Fig. 12 shows the doubling rate of 3 countries.

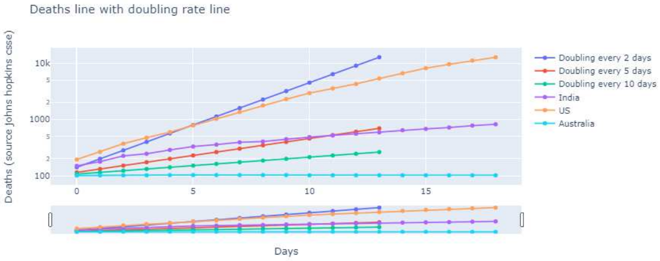


Fig. 12. Deaths Doubling Rate.

And below Fig.13 shows the similar trend with respect to recovered cases.

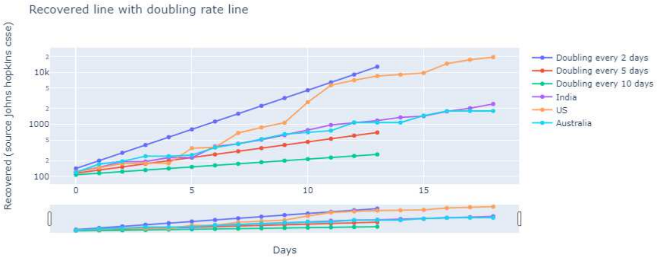
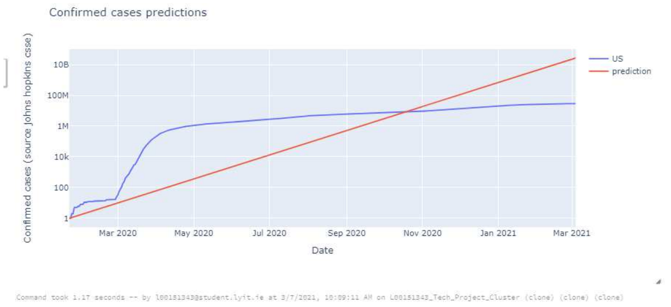


Fig. 13. Recovered Cases Doubling Rate.

Machine Leaning and Predictions:

For the purpose of predictions in the trends, we are going to use Supervised learning techniques [2]. At the same time we will fit a predictor line so that we can predict the values using the fitted line along the trend. Our target vector in this case is any of the columns that represent the values of the confirmed cases. We would like to predict the values in this target vector. Input vector represent the dates so that we can predict the values for a certain date.

Below Fig.14 shows how predictor line is fitted with the trend and



```
Command took 1.17 seconds -- by l00151343@student.lyit.ie at 3/7/2021, 10:09:11 AM on L00151343_Tech_Project_Cluster (clone) (clone) (clone)
```

Fig. 14. Predicting the confirmed cases.

In our case of the data that we are processing here really requires a prediction based on the dates so it is recommended to use the best fitted slope line for the prediction.

At last we are going to find a correlation between the confirmed cases, deaths and recovered cases for a specific country. The values achieved shows that these fields are highly correlated.

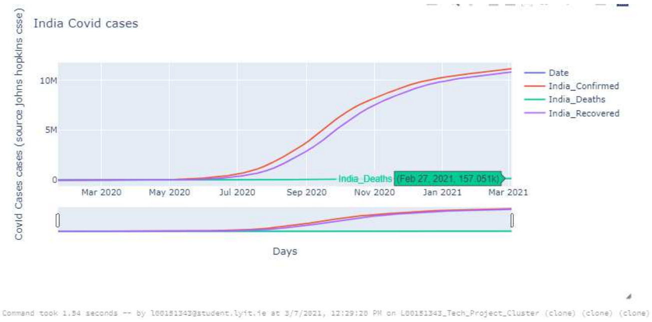
TABLE I
CORR VALUES CONFIRMED, DEATHS AND RECOVERED CASES

Num	Confirmed case	Deaths	recovered Case
1.	1.0	0.99733438582355 37	0.9976358301078774
2.	0.99733438582355 37	1.0	0.9906137784362529
3.	0.9976358301078774	0.9906137784362529	1.0

V. RESULTS AND ANALYSIS

Different approaches to analyse the covid data has provided insight into how the virus has travelled in community. Different graphs for counties provide which countries has been impacted in chronological fashion. As we reached the final stage, we can visualize that the trend of confirmed cases and recovered cases are quite closer on the plot. Fortunately Death trends show that we have high rate of recovery in our community. Still we need to adhere to the provided guidelines to until we have vaccination in place and restrict the spread of the virus in the community.

Below graphs shows the covid cases trends for a specific county to understand the recovery rates.



```
Command took 1.54 seconds -- by l00151343@student.lyit.ie at 3/7/2021, 12:29:20 PM on L00151343_Tech_Project_Cluster (clone) (clone) (clone)
```

Fig. 15. Covid cases.

VI. FUTURE WORK:

This work has been performed on a very simple data which revolves around the trend analysis. However there are lot of covid data is available today and we can explore it from many different point of view. Few that can be taken forward are:

- Covid data can be explored for a particular country or region to predict the infected numbers and how many requires medical intervention. Based on these insights governments will have prior readiness on its healthcare capacity to handle the situation. Applying predictive models also provides a forecasting on the number of caregivers required in a particular region.
- A proper Covid data when analysed with Big Data tools and AI can help in the diagnosis process.
- Not only in the healthcare domain, data analysis can help us to maintain proper demand supply in certain locality based on the predictions.
- With the help of social network analytic, contact tracking will be very easy which is a really tough task when in covid scenario it is mandatory to know the other infected people.

REFERENCES

[1] Q. Pham, D. C. Nguyen, T. Huynh-The, W. Hwang and P. N. Pathirana, ”Artificial Intelligence (AI) and Big Data for Coronavirus (COVID-19) Pandemic: A Survey on the State-of-the-Arts,” in IEEE Access, vol. 8, pp. 130820-130839, 2020, doi: 10.1109/ACCESS.2020.3009328.

[2] K Eremenko and H de Ponteves. Machine learning az: Hands-on python & r in data science. United States of America: Udemy. Retrieved May, 10:2019, 2019.

[3] Puhani, P.A., 2020. France and Germany exceed Italy, South Korea and Japan in temperature-adjusted Corona proliferation: a quick and dirty Sunday morning analysis (No. 487). GLO Discussion Paper.