Covid Analysis on USA : PCR Testing, Cases and Death, Hospital Utilisation, Reimbursement Claims

Anurag Ratnaparkhe
Department of Computing
National College of Ireland
Dublin, Ireland
x19229992@student.ncirl.ie

Dhanshree Bauskar
Department of Computing
National College of Ireland
Dublin, Ireland
x19230460@student.ncirl.ie

Dhwani Dharmesh Hingu
Department of Computing
National College of Ireland
Dublin, Ireland
x19216742@student.ncirl.ie

Mardwin Alejandro Cardenas Rodriguez
Department of Computing
National College of Ireland
Dublin, Ireland
x20144237@student.ncirl.ie

***Abstract*—The outbreak of COVID-19 disease started from one country and affected whole world in no time. COVID-19 is an contagious disease caused by newly discovered Coronavirus. The Coronavirus pandemic has affected our lives and economy and nearly every corner of the globe. Therefore, it has become important research area since it started spreading around the world. It has infected approximately 148,464,599 people in the world out of which 32,871,019 people are from United States of America followed by India with number of cases 17,625,735.After heart disease and cancer COVID-19 had become the third leading cause of death in USA in 2020, also more Americans have died due to COVID-19 compared to both world wars. Approximately there were 32 million confirmed cases and 572,000 confirmed deaths. It was observed that the first case in USA was detected in North America [1]. People with strong immunity power have been recovered from the disease, with no guarantee of getting infected again. It is been observed that people older than 50 years or underlying medical problems like High blood Pressure,Obesity,Diabetes, Chronic Respiratory disease are more likely to get infected easily and develop serious illness. This paper emphasizes on four distinct datasets like PCR Testing, New Cases and Death by state, Hospital Utilisation, State-wise Reimbursement which are inter-linked by state and the time period.**

I. Introduction

COVID-19 has impacted the world in a drastic way, initially misunderstood for a nominal virus like flu, has affected the whole wide world. Its been worst for the world economy, as its almost same in magnitude as the great recession back in 2008.The Major aim of this project is to analyze the effects of Covid on one of the most powerful and influential country in the world. The US is considered as the most advanced and powerful country in the world(both healthcare and infrastructure wise), and yet it is one of most impacted by Covid which directly affects the economy and balance of whole world. So in a sense we also evaluate the situation in general by analysing the impact of Covid on US. The analysis contains the PCR testing of unwell patients, also what is the amount of new cases getting reported everyday along with casualties if any if the patients are tested positive and are unwell, how many of them are hospitalized and what is recovery rate from Covid, and also is there any Relationship between critical staffing shortage which has been observed according to the data, and health recovery of the patients. How many amount of reimbursements have been paid by the insurance companies for the insured patients in terms of claims paid for testing then treatment from the disease and vaccination to cure it.

II. Related Work

The main purpose of this paper is to Implement advanced forecast models that could assist governments and health organizations to obtain useful future forecasts, this would aid governments and health organizations in collaborating and guiding to prevent the spread virus [2]. It is specified that The United States has been named the world's most impacted nation in general aspects, so the main purpose in our paper is to make a deeper analyst about the impact that covid has had in each state including daily total death, daily positive cases, hospital capacity impact and the payments that has made by insurance companies.

This paper it is based on study made just in Massachusetts and compares the impact of PCR testing on deaths, Infections, and hospitalisation over 180 days in four different testing strategies: Hospitalized (patients with critical symptoms), Symptomatic (PCR for any symptoms), Symptomatic + asymptomatic-once (Symptomatic and one-time PCR) and Symptomatic + asymptomatic - monthly (Symptomatic with monthly re-testing). The conclusion of this article is the infections, deaths, and hospitalizations would be reduced if PCR testing were extended to asymptomatic individuals [3]. This helps us on our study to create the relation between the same factor, as a connection can been observed between the total PCR testing applied and the total death and the total patients hospitalised in all US states.

This paper analyses that within three months of outbreak of covid-19 how many publications were written scientometrics approach. They analyzed that most publication were from journals and especially peer-reviewed article as well as publications came from medical background. It was worth noticing that most of the contributors came from three countries – China, USA and UK. This paper was helpful for this project to get some interesting insights about covid19 publications. One limitation of this paper was observed that they did not analyse about cases and deaths due to covid19 [4].

This paper analyses impact of covid19 on flight network. This paper demonstrates how many worldwide airports are and how flights were operated before and during COVID-19 also how many flights stopped operating or decreased after government announcements of travel restrictions. The paper was helpful for this project to analyse how covid19 cases were observed in USA due to high number of travel history. An interesting insight about distribution of number of airports and number of flights per area and country in form of donut chart was helpful in this project to analyse that USA had 33% of world's airport and 10.7% of the international flights were from USA from January 1 to April 30 in 2020. There was one limitation in this paper that they did not analyse about deaths in USA after tremendous increase of cases due to travel history [5].

This study computes the relationship between COVID-19 deaths and COVID-19 hospitalizations using actual data. The computation was done on the basis of current ICU patients and non-ICU patients. The study was conducted on total of 1056 observations across 23 states of USA. The author determined that the usage of ICU beds were higher than the non-ICU beds which was significantly associated with increase of overall COVID-19 deaths. The study was limited to COVID-19 mortality and exclusive of the states which expanded their healthcare system in COVID-19 era. In this analysis, hospitalization data is merged with three other datasets in order to perform a deep analysis about the COVID-19 statistics of USA states so that this paper can help to take further action involving battling with this epidemic [6].

Even though USA is considered to be one of the most powerful and developed country, most of the people in country still lack health insurance (around 27.5 million as of 2018), these are the people who belongs to lower socioeconomic status and ethnic minorities. Most of the employed people are given health insurance facility from their organisation, and the people not having insurance faces financial issues mainly when disease like COVID-19 have spread all over the world. In this project, with the help of medical insurance reimbursement data it is been analysed that major claims were made from the people of California, Texas and New York as people in these cities are more probable to have an employer who sponsors Health Insurance [7].

The paper analyses the effects of social distancing in USA county wise, The Data-set for the purpose is being extracted by different sources by tracking the mobile devices which is done by Safegraph's database and then merged together to create a final data-set.The Analysis is done by calculating the time for which a mobile device is stationary at a place, and the time it is travelling and then comparing with other device location.One similar finding which has been showed by this paper is that the highest COVID-19 affected states according to this analysis also turns out to be California and Texas and New York which also is the case in our paper. The strong point of this analysis is that it clearly infers that by following social distancing guidelines the cases had a significant drop. One Limitation which this paper does not cover is that what if a person is not keeping phone with him/her always and the people whose devices are not being tracked [8].

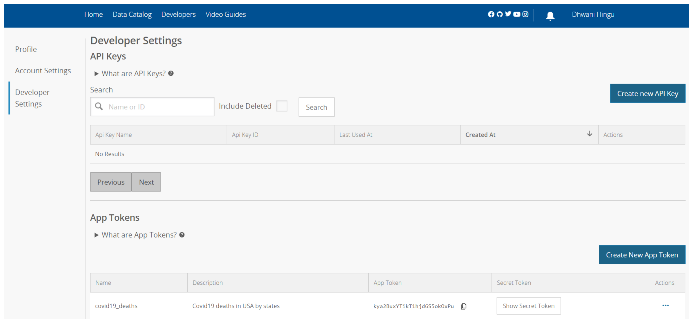
The analysis is done on the spread of COVID-19 in India states on the basis of climatic influences data is collected from Indian government climate website and data about COVID-19 cases is collected from a Indian origin website which is keeping track of COVID-19 in India. First, the states are classified according to their climatic conditions, and 6 different climatic zones were defined. Using the GAM model, an attempt was made to relate the number of infected cases with all the geographical variables. The final conclusion is made that the the climatic conditions does not play any significant role in the spread of COVID-19.One Limitation of this analysis is that it does not consider evaluating the population of the state with its climatic condition which can be a factor in the analysis [9].

III. Methodology

The analysis consists of four individual datasets which are merged together to draw some useful insights about the impact of Covid-19 on USA.

A. Dataset Description and Gathering

1) Dataset 1 - Diagnostic Laboratory Testing: This time series dataset [10] is a collection of the results of viral COVID-19 testing from over 1,000 US laboratories and testing sites. In compliance with relevant state or local laws, data is submitted to state and jurisdictional health departments. The information that is present in this dataset is date, state, fema region, overall PCR results, new results reported every day and the total of results reported. The PCR testing that are found in this dataset are: Positive, Negative and Inconclusive test results, so this allows to have a general picture about the total PCR testing impact by each state. The principal sources of this dataset are: Commercial Laboratories, Electronic Laboratory, Hospital Labs and State Public Health Labs. This data set has been chosen base on the important that is implicated to confirm if a person has been tested even positive or negative. In the related works, it clearly specified the value of PCR testing to reduce the deaths and the high hospital demand caused by COVID- 19.



Developer Settings

API Keys

▶ What are API Keys?

Search

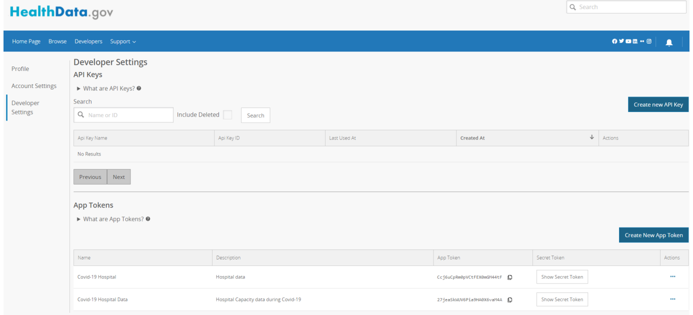
Api Key Name	Api Key ID	Last Used At	Created At ↓	Actions
No Results				

App Tokens

▶ What are App Tokens?

Name	Description	App Token	Secret Token	Actions
covid19_deaths	Covid19 deaths in USA by states	kya2BuxYTikT1hjd6S5okOxPu	Show Secret Token	···

Fig. 1. API from CDC



Developer Settings

API Keys

▶ What are API Keys?

Search

Api Key Name	Api Key ID	Last Used At	Created At ↓	Actions
No Results				

App Tokens

▶ What are App Tokens?

Name	Description	App Token	Secret Token	Actions
Covid-19 Hospital	Hospital data	Ccj6uCpRm0pVCtFEX0mGM44tF	Show Secret Token	···
Covid-19 Hospital Data	Hospital Capacity data during Covid-19	27jeaSkWUV6Pia9HA0X6vaM4A	Show Secret Token	···

Fig. 2. API from Health Data

2) Dataset 2 - Cases and Deaths by State : The dataset was fetched from Centers for Disease Control and Prevention(CDC) [11] through Socrata Open Data API programmatically. By using API endpoint only one thousand random records were fetched and the dataset requirement for this project was minimum five thousand records, so account was created on CDC developers section and an app token was created in order to get time series data which consist of more than one thousand data and the data is updated every day. Sodapy is a python client for Socrata open data API to connect to the CDC dataset. The dataset is about COVID-19 cases and deaths in United States and is it time series data where data is updated daily and some of the features of dataset are submission_date, state, tot_cases, conf_cases, prob_cases, new_case, pnew_case, tot_death, conf_death, prob_death, new_death, pnew_death, created_at, consent_cases, consent_deaths. COVID-19 illnesses, hospitalization, deaths, cases etc are tracked by CDC so the data is very accurate. COVID-19 causes mild illness, symptoms might not show up immediately, thus there are delay in reporting and testing, many a times a person is detected positive for COVID-19 and the person does not have any symptoms on the other hand the person has symptoms and the person is detected negative, so it is not possible to keep the exact count. Therefore there is difference in each state and territory to report cases and deaths. On weekends and holidays there Is generally less reporting. There is fluctuation in reporting for number of cases each day as heath departments may update case data overtime when they receive more complete and accurate information. The rational for choosing this dataset was because the dataset comes from official source which is CDC, it is open source, it has API to scrap data, consist of state column which was helpful for joining other datasets for group analysis. One of the most important aspect of choosing this dataset was that in consisted complete state wise and date wise data about how many new cases got reported, confirm cases, new deaths and confirm deaths also how many probable cases and deaths are reported.

3) Dataset 3 - Hospital Utilization: The data is a collection of state wise Hospital utilization in a timeseries format published by US Department of Health & Human Services [12]. The data is obtained from the major sources: HHS Tele Tracking and state health departments. It gets updated regularly and provide the latest values. Each row in the dataset contains the collection of daily report stating the staffing shortage in the hospital, patients hospitalized suspected and confirmed COVID-19, number of beds used by the patients. As it is a time-series data every day it gets updated, for the scope of analysis date and limit of data is given in API request, so the retrieved data will be from 1 December 2020 with limit of 30k records. The reason behind choosing this dataset is it is available on official source which is Health Data, it is open source and has option to get data from API also the dataset contains the column through which it can be connected with other datasets in the project.

4) Dataset 4 - Reimbursement: The Dataset is related to impact of COVID-19 on the states of USA. The Dataset is extracted from the official government operated website [13] of US which keeps track of healthcare records. The raw data file is available to download in XMl semi-structured format. The Dataset initially contains 8 distinct columns namely, 'id', 'provider name', 'state', 'city', 'claims paid for testing', 'claims paid for treatment', 'claims paid for vaccine', 'geocoded column' and 29146 rows. Xpath notation is used to navigate to the the child nodes which contains the useful information. This is done because there is a complexity in the raw file where there are two child nodes of same name 'row' are defined under root node with the hierarchy as follows Root - Row(1st child)-Row(2nd child)-Data to be extracted, so in order to properly navigate to the correct child node Xpath notation in combination with iterfind method is used in order to extract the data from raw XML file and store in empty lists. The main aim of the dataset is to keep record of what amount of claims have been paid for the testing, treatment and vaccination of the covid respectively on the basis of city and state. This particular dataset is selected for analysis because, it shows proper data for the amount of claims paid by different insurance companies in different states of US and when merged with other datasets which have state wise data , interesting patterns are found.

B. Data Preprocessing

1) Dataset 1 : Diagnostic Laboratory Testing: For Covid Testing dataset, data was obtained via API from HealthData.gov as a JSON format. MongoDB was chosen as the data base to store it as MongoDB allows to work with unstructured data format, meaning that the structure do not need to be defined first. After the unstructured data has been fetched back from MongoDB it was transformed it into pandas data frame in order to proceed with the cleaning process, the presence of NA value were checked, but there were not any wrong value, the unwanted columns from the dataset have been dropped, for example: the state_fips and geocoded_state. Original dataset used to have three different records for the same day, each one depending on the result of the test. To reduce the number of rows, data was split into tree different dataset base on their respective result (Positive, Negative, Inconclusive). The last tree tables were merge day wise to have just one record per day, this reduced the total amount of rows tree times and at the same time allows to have a daily record with all the results that were measured. Finally, the appropriate data type for each column were assigned because taking care of data types is very important to avoid issue while data is processing or merging. Once the final data set was clean and transformed into the final structured data set, it was pushed to PostgreSQL as our main data warehouse.

2) Dataset 2 : Cases and Deaths by State : For data preprocessing the data was fetched from MongoDB and stored as Dataframe, after creating a DataFrame the data was filtered with date from 1st of December 2020 to current date because wanted to study and analyse interesting trends of cases and deaths in each state considering last 5-6months data. For data cleaning, the data was checked programmatically whether it consists of null/missing values as well as if there are any duplicate entries. It was observed that there was almost 88% of missing values, these values were missing because as per CDC these counts are continuously revised and updated on daily basis as well as some jurisdictions do not report daily so there are blanks in the report. Dropping null values was not a good option because as the data is time series data and while merging data for group analysis based on date and state it would have affected overall analysis. Also, imputing data statistically was not a good option as it would have created a bias approach for individual analysis as well as for group analysis. Thus, the data was filled with zero indicating that particular information is missing at the moment as the data is time series data and the data keeps updating on daily basis. Most of the numeric fields were in object data type so these fields were converted to integer data type.

A new column was created named delay_reporting by subtracting created_at and submission_date, because created_at column indicates the date and time the report was created and submission_date means the report was submitted on that date. So, by subtracting two columns, by how many days the reporting of cases and deaths was delayed for given states. Although for most of the cases the report was submitted on same day or next day but on the other hand it was observed that there was delay in reporting for more than 2 to 141 days

3) Dataset 3 :Hospital Utilization: The data for Hospital utilization is been extracted through an API call from the site [] with the help of API key. Raw data was in semi structured JSON format. Mongo DB was used to store semi-stuctured data and it can be fetched in the form of dataframe using Pandas Python library. As the raw data had 61 columns, only the selected columns like 'state', 'date', 'critical_staffing_shortage_today_yes', 'inpatient_beds_used_covid', 'inpatient_beds_used', 'total_adult_patients_hospitalized_confirmed_and_suspected_covid', 'percent_of_inpatients_with_covid', which will be helpful to draw insights are chosen and stored in new dataframe. The process of data cleaning was done on the new dataframe, it includes finding null values and missing values and if present they need to be removed or replaced with appropriate data, but in this data no null values or missing values were present. Changing the data types of features is an important part to proceed futther with database connections. Every feature should be in appropriate datatype like ID should be integer, Date should be in datetime format. The cleaned and structured data is now pushed to postgreSQL for connecting with other datasets.

4) Dataset 4 : Reimbursement: The dataset was downloaded in the form of semi-structured data and was stored in MongoDB. The primary objective of this project is to analyse the impact of covid on different states of US thus only including the columns which is going to help to accomplish the final goal. As a result the id, geocoded_column were initially dropped from being included in the further analysis. also null values were checked and only one row was identified as null and thus was dropped. The raw dataset also included some special characters like'$' and '-' etc, which were cleaned. Lastly the columns which represented claims paid are essentially payments and were converted to int datatype for further processing.

Python: We chose python because it has uncomplicated and understandable language like English and as soon as it is written the code get executed that mean it's an interpreter framework. Python can be run on various platforms (Windows, Mac, Linux, Raspberry Pi, etc.). The foremost import aspect to choose python was that it supports packages and modules which make developers task easy as one can reuse code and encourages program modularity.

MongoDB: As we fetched time series data through API, the data was in Json format so it was easy for us to load the data into MongoDB. MongoDB is open source and easy for developers to learn. Also we used PyMongo to interact with MongoDB. Ad hoc queries, indexing, and real-time aggregation are all useful tools for accessing and analysing data which make developers task easy thus we chose MongoDB.

PostgreSQL: Several features, such as native partitioning, parallel query, support for international data wrappers, powerful JSON features, streaming and logical replication, and the availability of several open source tools for HA, backups, and monitoring, have made PostgreSQL popular among developers. Its free and open source object-relational database system as well as it highly reliable. Also if you don't like any characteristic of framework you can change and customize as per the needs.

- **Diagrams highlighting the data gathering, processing and analysis flow will be useful here.**

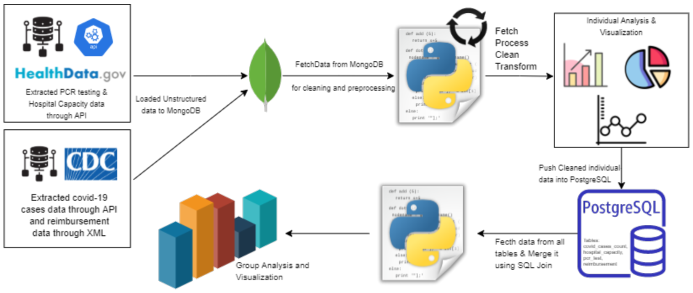

HealthData.gov
Extracted PCR testing & Hospital Capacity data through API
CDC
Extracted covid-19 cases data through API and reimbursement data through XML
Loaded Unstructured data to MongoDB
FetchData from MongoDB
for cleaning and preprocessing
Fetch
Process
Clean
Transform
Individual Analysis & Visualization
Push Cleaned individual data into PostgreSQL
PostgreSQL
Tables:
covid_cases_count,
hospital_capacity,
pcr_test,
reimbursement
Fecth data from all tables & Merge it using SQL Join
Group Analysis and Visualization

Fig. 3. Process Flow Diagram

The fig.3 depicts that datasets were fetched through API from Heathdata.gov and Centers for Disease Control and Prevention(CDC), also it was open source. One of the dataset was fetched through XML. After data gathering, individual databases and collections were made programmatically as well as interacted to MongoDB using PyMongo library of python, and the data was loaded to MongoDB. Later for data cleaning, pre-processing and transformation the data was extracted from MongoDB and stored as Pandas Dataframe because Dataframe is a structure that holds two-dimensional information. Also Its very convenient to use dataframe because its similar to excel and widely used for data analysis and machine learning. Once the data got cleaned, wrangled, transformed, individual analysis was done to show case and highlight interesting trends through visualizations by plotting graphs. The cleaned and transformed data was than programmatically pushed to postgreSql. The data was then fetched from postgreSql and merged programmatically by using SQL join so that merged analysis can me done and based on the same, some noteworthy visualization was demonstrated.

IV. Results

The fig.4 is a scatter plot which shows new cases in each state of USA. Highest number of new cases are observed in California(CA) state that is approximately 53k and it was reported on 16th December 2020. There was false reporting done on 8th of January 2021 in New Jersey(NJ) almost around 9k false new cases were reported. So, on the next day 9th of January 2021 these 9k records were marked as negative indicating as false reporting. Although these 9k records were subtracted from total number of cases to make it adjusted. The Republic of the Marshall Islands(RMI), America Samoa(AS), Palau(PW), Federated States of Micronesia(FSM) have zero cases, because these are USA islands/territories and they have relatively less population also during COVID-19 there was no travel history to these places hence there are no cases.

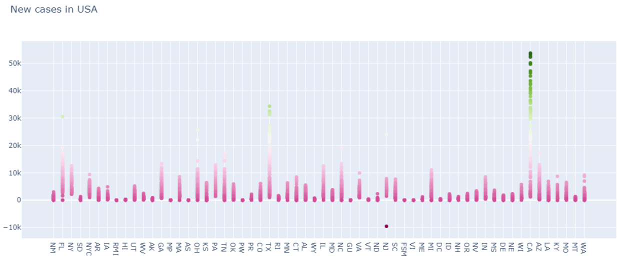


Fig. 4. New Cases in USA

On the other hand, fig.4 illustrates that following states had high number of cases:
California(CA): approx. 53k new cases
Texas(TX): approx. 34k new cases
Florida(FL): approx. 30k new cases
Ohio(OH): approx. 25k new cases
The above states are most populated that's why there were high number of chances that people got detected to COVID-19. Also there was a lot of travel history during COVID-19 in these states as these states consists of large number of people traveling for business or students traveling to and from there home countries.

In fig.4 its been observed that California had highest number of cases, so for deep analysis another visualization in fig.5 was done considering only California state
The data was filtered from 1st December 2020, so 70 new deaths were reported and by the 26th December 2020 deaths had got decreased to 36 but than next day 27th the deaths got increased and by the end of December 428 deaths were reported.
By 1st of January 2021 585 new deaths got reported and on 19th January a fall of deaths were suspected. In the graph it is highlighted that on 22nd January there were highest number of new deaths reported which were 764, two day later new deaths rate got half followed by day after the deaths again got spiked again to 737.
The line graph was considered as it depicts interesting trend as there is fluctuation in new deaths each day in California. Substantial fall of deaths have been witnessed in February and March, By April it depicts that there are considerably lowest number of deaths and the situation in California are in under control as people are getting vaccinated.

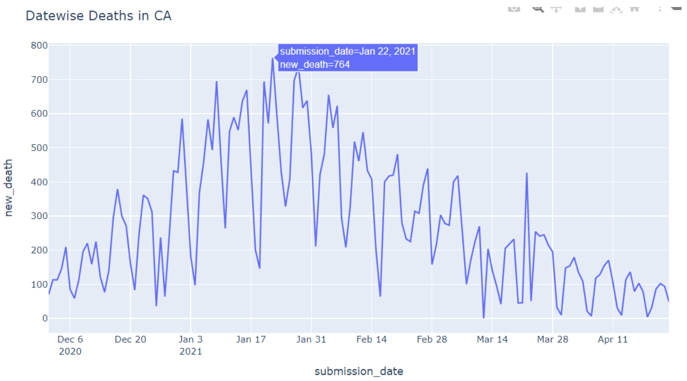


Fig. 5. Deaths in California

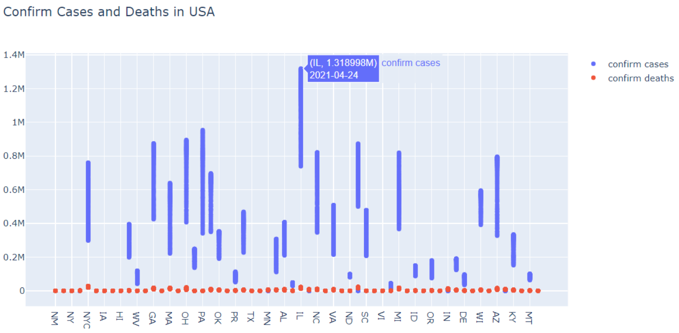


Fig. 6. Confirmed cases vs deaths in CA

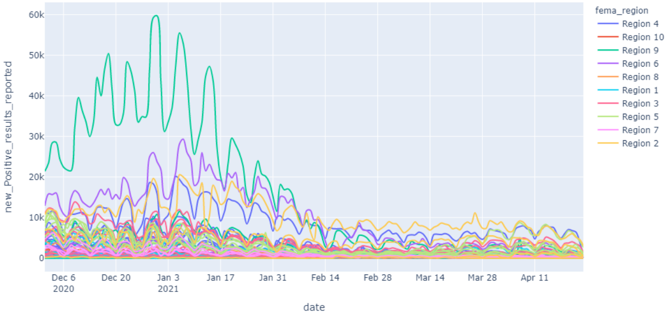


Fig. 7. New positive results reported

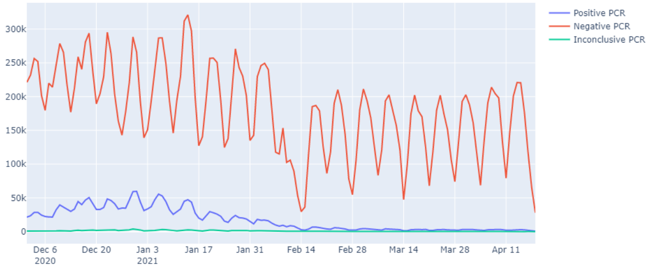


Fig. 8. PCR TESTING RESULTS IN CALIFORNIA

The fig.6 graph demonstrates number of confirm cases as blue dots and confirm deaths as red dots in each state of USA. The highest number of confirm cases have been found in IL(Illinois) which is approx. one million on 24th April 2021, followed by Pennsylvania had 953k confirm cases on 24th April 2021. On the other hand states like NM,NY,AR,IA,PW,DC etc few more had zero confirm cases. Compared to confirm cases, there are less number of confirm deaths. Thus, this indicates that less number of people are dying which is considerably good. State NYC had approx. 27k confirm deaths on 24th April 2021, which is highest number of confirm deaths compared to other states.

The fig. 7 is a representation of the positive COVID tests that were conducted by every region in the United States in the last five months, at the same time each region is divided by all the states, so in this way we can observe that the state with the highest number of positive cases is California which belongs to the 9th. region.

In figure 7 has been spotted that California as the state has been more impacted by COVID in the last 5 month,so in figure 8 we proceed to make an specific analysis about the total amount of COVID-19 tests that has been apply and how many of them were positive, negative or inconclusive each day. We can see in the graph there is substantial decrease in the positive test made after January, nonetheless, the negative PCR test amount remains at the same levels, so people were still getting tested but there were less positive case.

Figure 9 depicts the number of patients hospitalised who were suspected and confirmed COVID-19 in the United states. It can be seen that the major number of patients hospitalised were from California, Texas and New York, that means these were the most affected states of USA.

Figure 10, the Pie chart represents the percentage of staffing shortage in the respective states of USA. The percentage of staffing shortage means the number of hospital staff member is less than the patients admitted or the beds occupied in the hospital. In the figure it can be seen that Texas had the higher shortage rate with 14.8% followed by California with 13.3%

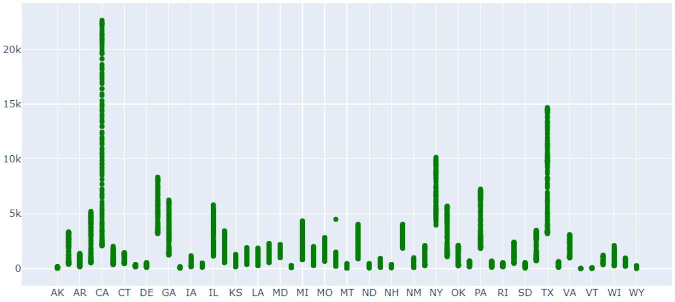


Fig. 9. Patient Hospitalised

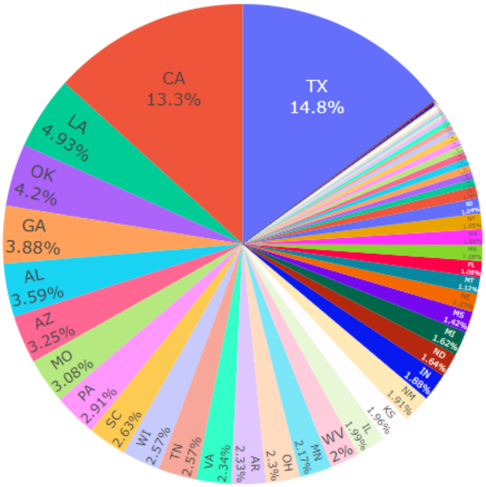


Fig. 10. Hospital Staffing Shortage

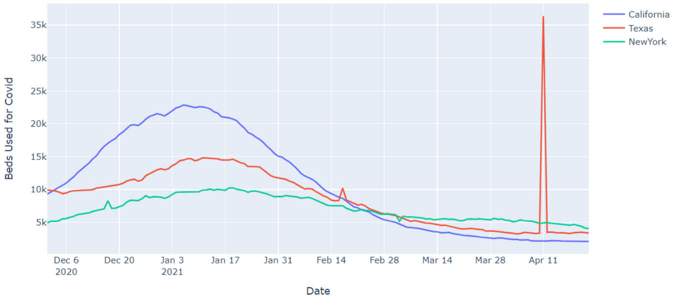


Fig. 11. Usage of Beds for COVID-19

and LA with 4.93%.

Figure 11 illustrates variation of Hospital beds used for COVID-19 patients in three major affected states California, Texas and New York. In the start of December 2020 New York had 5k beds for COVID-19 which is comparatively less than California and Texas, this continues till mid of February and around 20-25th of February almost 8k beds are occupied in these cities. Texas gets highly affected on 11th April and reaches beyond 35k utilisation of beds for COVID-19 whereas California and New York are under 10k.

In order to better understand the nature and distribution of different values in the Dataset a Donut Pie chart and a scatter plot is plotted, plotly is used as a data visualization package, because it provides easy to use and quick to process plots.Fig 12 which is a scatter plot illustrates the distribution of claims paid by various insurance companies for COVID-19 testing in USD as per each state. The figure illustrates that the state of Texas and California are the top 2 states where claims are made. Fig 13 is Donut pi chart which shows the distribution of claims paid for the purpose of COVID-19 treatment here also CA and Texas are the top 2 states which adds up as they are also the states where maximum claims are made for testing,

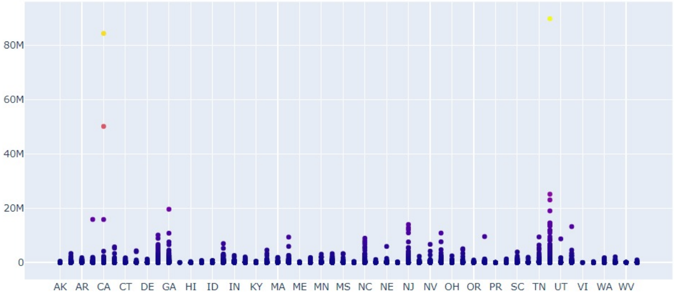


Fig. 12. Claims paid for testing

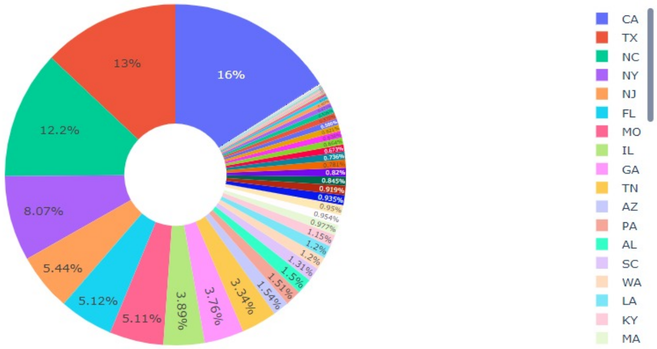


Fig. 13. Claims paid for treatment in USD

Looking at the data its been clear that Texas and CA are among the top states which are most impacted by COVID-19.

A. Group Visualisation

The group graph in figure.14 illustrates relationship between total confirm cases & total positive results reported by each state. As seen in the above graph that Illinois (IL) had approx. 1.31M confirm cases as well as there were approx. 1.30M positive results reported. Thus, this clearly demonstrates that there were high number of new cases reported in Illinois also approximately equal amount of PCR testing was conducted. On the other hand its observed that Maine(ME) state had

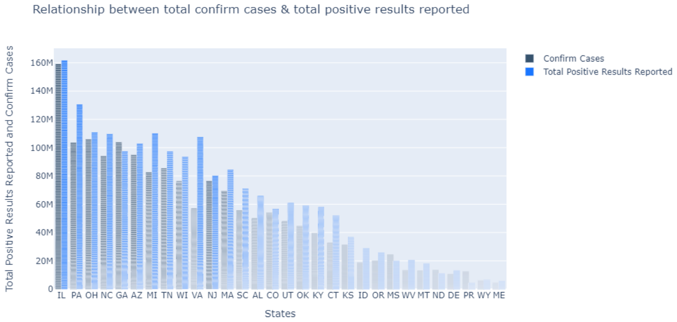


Fig. 14. Confirmed Positive cases

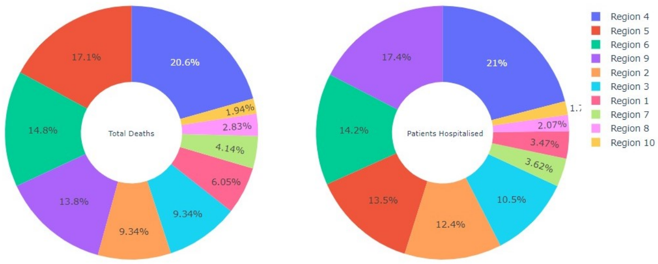


Fig. 15. Patients hospitalised vs death

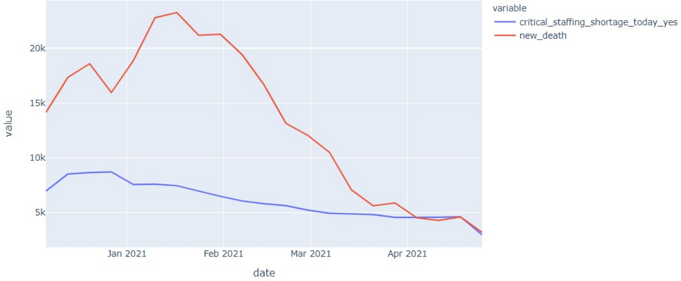


Fig. 16. Staffing shortage against new death analysis

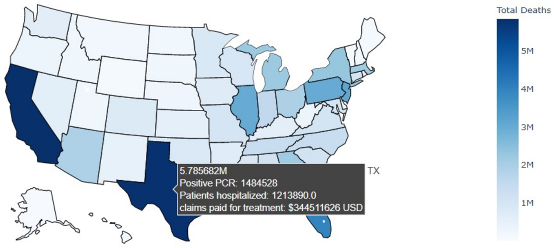

Total Deaths
5M
4M
3M
2M
1M
5.785682M
Positive PCR: 1484528
Patients hospitalized: 1213890.0
claims paid for treatment: $344511626 USD
TX

Fig. 17. Most impacted US states by COVID-19

lowest number of positive tested cases and lately confirmed by ME state.

In figure. 15 Donut chart-1(left-side) shows occurance of total number of deaths in the Fema-regions and the second donut depicts the number of people hospitalized in Fema-regions. Analysing donut 1, region 4 has highest death of 20.6% followed by region 5 and region 6. By studying both the charts we can calculate the percentage of people recovered from COVID-19, by subtracting number of Deaths by number of people hospitalised in the particular regions of USA.

Fig. 16 is a comparative line graph which answers the question is there a impact of critical staffing shortage on the new deaths. The blue line(or the line with lower values) represents the critical staffing shortage in the hospital and the orange line (or the line with higher values) shows the new deaths, after plotting the lines together on the same scale that is during the period of December to April, the plot gives an insight that there indeed is a Relationship between the two, turns out as the slope of staffing shortage goes up by some unit the new deaths also increases and in mid Feb when the staffing shortage is on decline or in other words when hospitals start to stabilize the number of new deaths also declines exponentially.It is a great insight which clearly shows the importance of good healthcare infrastructure and medical workers on the impact of handling the COVID-19 situation.

The Figure 17 has been created to summarize the effect of COVID-19 has had in USA and have a general visualization of COVID-19 impact by state, States has been ranking in scale of blues meaning the darkest blue represent the state that has had the highest number of the total death in the last 5 months. Each state also includes a label with a more specific information of the totals of positive PCT testing results, total of patients confirmed with COVID-19 that were hospitalized and the total amount of claim paid by insurance companies. Los Angeles with more than 5.8 Million deaths in the last months is one of the most impacted state in US just follow for by Texas with 5.7 Million deaths, and if the people that has been hospitalized and get tested positive is compered too, it can be notice Texas has also lower values that Los Angeles. Also, companies in Los Angeles had paid around $1 M more than in Texas for COVID-19 claims.

V. Conclusion and Future Work

This paper introduced Data Analytics approach to figure out Covid-19 impacted areas of the country with the help of visualisation. As with this analysis it can be inferred that the most affected states of USA by Covid-19 were California, New York, Texas and Illions. With the combination of distinct datasets, it can be observed that hospital staffing shortage was affecting the mortality rate. This study can be extended by adding vaccination data as even when it has been more than a year that the Covid has impact the world, vaccination programs are in early stages, so there are not too many statistics about vaccination, so vaccination data could be analysed in the future to see if the vaccination is a key point to reduce the cases and deaths. A deep analysis could be done including the main action taken by governments and see how this action helps to reduce the spread of covid cases, in this way the action could be evaluated if they were good to stop the covid spreading or not. Lockdown is one of the main actions that many countries had adopted to try to reduce the spread of covid 19, but how effective is this measure.? This is another factor which could be analysed as we could had compared against our time series analysis. Some machine learning models could have been implemented so as to find more deeper relationship between different variables.

References

[1] “Coronavirus,” World Health Organization. [Online]. Available: https://www.who.int/health-topics/coronavirus. [Accessed: 28-Apr-2021].

[2] A. K. Chandani, “Worldwide differences of Covid-19 on cases and deaths using time series forecasting models,” Worldwide differences of Covid-19 on cases and deaths using time series forecasting models - NORMA@NCI Library, 01-Jan-1970. [Online]. Available: http://norma.ncirl.ie/4434/. [Accessed: 28-Apr-2021].

[3] A. M. Neilan, E. Losina, A. C. Bangs, C. Flanagan, C. Panella, G. E. Eskibozkurt, A. Mohareb, E. P. Hyle, J. A. Scott, M. C. Weinstein, M. J. Siedner, K. P. Reddy, G. Harling, K. A. Freedberg, F. M. Shebl, P. Kazemian, and A. L. Ciaranello, “Clinical Impact, Costs, and Cost-Effectiveness of Expanded SARS-CoV-2 Testing in Massachusetts,” OUP Academic, 18-Sep-2020. [Online]. Available: https://academic.oup.com/cid/advance-article/doi/10.1093/cid/ciaa1418/5908298. [Accessed: 28-Apr-2021].

[4] B. Harsanto, ”The First-Three-Month Review of Research on Covid-19: A Scientometrics Analysis,” 2020 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC), 2020, pp. 1-6, doi: 10.1109/ICE/ITMC49519.2020.9198316.

[5] T. Suzumura et al., ”The Impact of COVID-19 on Flight Networks,” 2020 IEEE International Conference on Big Data (Big Data), 2020, pp. 2443-2452, doi: 10.1109/BigData50022.2020.9378218.

[6] P. Karaca-Mandic, S. Sen, A. Georgiou, Y. Zhu, and A. Basu, “Association of COVID-19-Related Hospital Use and Overall COVID-19 Mortality in the USA,” Journal of General Internal Medicine, 19-Aug-2020. [Online]. Available: https://link.springer.com/article/10.1007/s11606-020-06084-7. [Accessed: 28-Apr-2021].

[7] S. A. M. Khatana and P. W. Groeneveld, “Health Disparities and the Coronavirus Disease 2019 (COVID-19) Pandemic in the USA,” Journal of General Internal Medicine, 27-May-2020. [Online]. Available: https://link.springer.com/article/10.1007/s11606-020-05916-w. [Accessed: 28-Apr-2021].

[8] T. Banerjee and A. Nayak, “U.S. county level analysis to determine If social distancing slowed the spread of COVID-19,” Rev Panam Salud Publica;44, jun. 2020, 01-Jun-2020. [Online]. Available: https://iris.paho.org/handle/10665.2/52418. [Accessed: 28-Apr-2021].

[9] A. Gupta, S. Banerjee, and S. Das, “Significance of geographical factors to the COVID-19 outbreak in India,” Modeling Earth Systems and Environment, 17-Jun-2020. [Online]. Available: https://link.springer.com/article/10.1007/s40808-020-00838-2. [Accessed: 28-Apr-2021].

[10] U.S. Department of Health &amp; Human Services, “COVID-19 Diagnostic Laboratory Testing (PCR Testing) Time Series,” HealthData.gov, 28-Apr-2021. [Online]. Available: https://healthdata.gov/dataset/COVID-19-Diagnostic-Laboratory-Testing-PCR-Testing/j8mb-icvb?category=datasetamp;view_name=COVID-19-Diagnostic-Laboratory-Testing-PCR-Testing. [Accessed: 28-Apr-2021].

[11] “United States COVID-19 Cases and Deaths by State over Time,” Centers for Disease Control and Prevention. [Online]. Available: https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36. [Accessed: 28-Apr-2021].

[12] U.S. Department of Health amp; Human Services, “COVID-19 Reported Patient Impact and Hospital Capacity by State Timeseries,” HealthData.gov, 24-Apr-2021. [Online]. Available: https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/g62h-syeh. [Accessed: 28-Apr-2021].

[13] “Claims Reimbursement to Health Care Providers and Facilities for Testing, Treatment, and Vaccine Administration of the Uninsured,” Centers for Disease Control and Prevention. [Online]. Available: https://data.cdc.gov/Administrative/Claims-Reimbursement-to-Health-Care-Providers-and-/rksx-33p3. [Accessed: 28-Apr-2021].