# Analysis of the Pandemic, Mortality Rate and Insurances

**Mohammed Omar Khan(mok236@nyu.edu), Akshat Jain(aj3186) and Ankit Sati(as14128)**

Courant Institute of Mathematical Sciences, New York University

## ABSTRACT

Respiratory sickness Coronavirus Disease 2019 (COVID-19) quickly reached epidemic levels. Governments have had to take drastic measures to stop the spread of the COVID-19 disease due to the damage it has done to the world economy. Governments will be able to alter policy and prepare for the future if they can accurately foresee the effects of unlocking in the US. The temporal dynamics of COVID-19 cases were examined using the Autoregressive Integrated Moving Average (**ARIMA**) model, the Seasonal Autoregressive Integrated Moving Average (**SARIMA**) and we will show the error rate and forecast of Long Short-Term Memory (**LSTM**) using two different methods. We considered the number of confirmed COVID cases and provided the impact of the key risk factors for COVID cases. Based on a variety of factors, the reduction rates in different states differed. The sample of time series data was gathered between January 2020 and May 2022, and the analysis and prediction were carried out for three months following that date.

## 1 INTRODUCTION

The World Health Organization (WHO) claims "A new type of coronavirus called COVID-19 is contagious. This sickness was known as the "2019 novel coronavirus," or "2019-nCoV." Corona Virus was the abbreviation. The COVID-19 virus is a novel pathogen that is related to SARS and other common cold viruses." It results in symptoms that resemble those of the flu. Fever, cough, shortness of breath, pneumonia, and breathing issues are some of the signs and symptoms of influenza. It is a highly contagious illness that is spread by person-to-person contact as well as direct contact with respiratory droplets produced when an infected person coughs or sneezes. In Wuhan, Hubei Province, China, on December 29, 2019, the illness was first discovered (Li et al., 2020a). Since then, it has been unstoppable and has rapidly spread over the world. On February 21, 2020, the WHO declared COVID-19 a global health emergency following the reporting of 76,288 confirmed cases in China. In an effort to calm the situation, governments from all over the world tried lockdown, social isolation, sanitization, work from home, travel bans, and other measures. On the other side, people were becoming ill, and many nations were battling to maintain order. Governments are still having trouble solving the issue, despite several efforts.

The morbidity and mortality rates of COVID-19 were first unpredictable, particularly for young children and the elderly. The COVID-19 virus killed patients in Italy who were on average 81 years old and who had a history of smoking, diabetes, cancer, or cardiovascular disease [1]. Although it is clear that older people are more susceptible to the physical effects of SARS-CoV-2, this does not suggest that solely physical effects should be a cause for concern . Being restricted to one's home may be detrimental to one's mental, emotional, and physical health, as people all over the world are learning. Additionally, COVID-19 has exposed a critical vulnerability in current homecare policies for the elderly, which were expensive and hard to get before the outbreak. Contact with an infected individual can spread this virus, most frequently by microscopic droplets made when sneezing, coughing, or talking [2]. The timing was erratic in the beginning, and they knew nothing about the COVID-19 properties. Therefore, the distribution of COVID cases was significantly influenced by population density [6]. If a pandemic happens in the future, stricter regulations or implementations of distancing should be investigated in densely populated areas because population density has been recognized as a potential indicator of viral transmission.

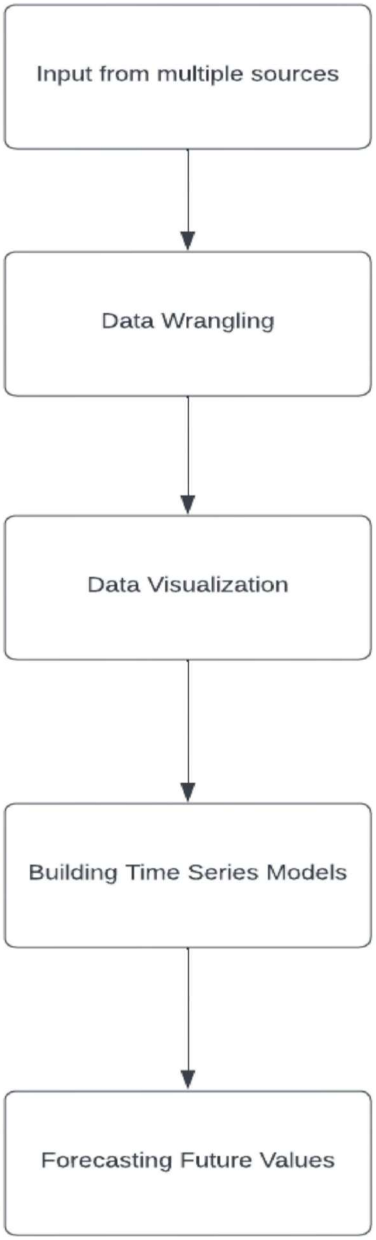
The weather was another factor in the disease's spread. According to the research [5], nearly 3 million reported cases (out of around 29 million tests conducted) of COVID-19 have occurred in places with temperatures between 3 and 17 °C and absolute humidity between | and 9 g/m3, or approximately 85% of cases recorded up to 1 May 2020. Similar to these hot and humid places, just 15%, or around 0.5 million cases, have been documented outside of these ranges (out of approximately 7 million tests performed). This shows that weather has an impact on COVID-19's global dissemination. Hospitalizations rose because of the general increase of COVID patients [4]. The daily trend of hospitalizations after state reopening was higher by 1.607 per 100000 population in the cross-sectional examination of COVID-19-telated hospitalizations and deaths across 47 US states between April 16 and July 31, 2020. Patients' underlying medical issues also had an impact on their immunity, which was another contributing factor to the disease's severity and the need for hospitalization.

Using time series analysis, we attempted to estimate the instances from May 1, 2022, to August 1, 2022, in this study. Predicting the covid-19 confirmed cases scenario with a rough number range between May 1 and August 1 of 2022 was the goal. Additionally, a model that would be excellent for forecasting the COVID-19 situation had to be developed. The ARIMA, SARIMA and LSTM Models were employed in this study to anticipate confirmed COVID-19 cases three months in advance. The time series analysis data from January 22, 2020, to May 1, 2022, served as the basis for this model.

In recent years, a variety of techniques have been employed in time interval prediction. They frequently incorporate machine learning algorithms, empirical models, and remote sensing techniques. However, machine learning models—which are most frequently used in artificial neural networks (ANN), such as multilayer perceptron neural networks, evolutionary ANN, generalized regression neural networks (GRNN), and backpropagation neural networks—are the most promising methods to predict forecast with their high accuracy (Feng et al., 2020). The alternative approach, known as time series analysis, is one that is frequently employed due to its high accuracy for small datasets.

## 2 METHODOLOGY

Figure | depicts the methods we used to carry out the studies. The procedures below are used to extract the data from various sources.


Input from multiple sources
Data Wrangling
Data Visualization
Building Time Series Models
Forecasting Future Values

For various visualizations, we combine data from diverse sources. These files provide a variety of information, including the number of COVID 19 cases and the number of hospitalizations in each state. The data wrangling stage is completed to aid with model training and visualization since both processes require a clean dataset to produce accurate results. Following this stage, we produced several visualizations before training the time series models.

## 2.1 DATASETS

We made use of several datasets. Their characteristics are shown in the table below.

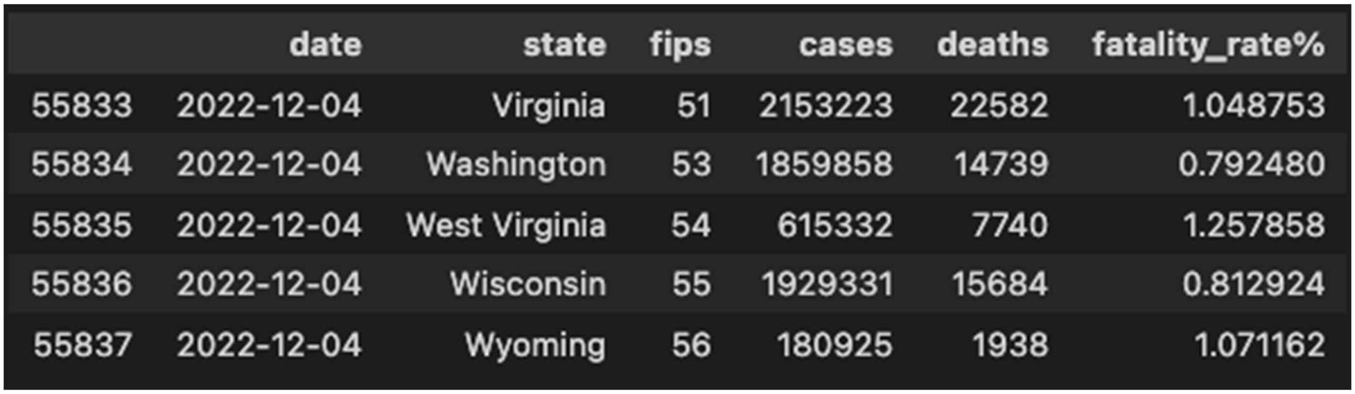
| Attributes | Attributes Description |
|---|---|
| State | state All the states in US |
| Date | date From 2020-01-01 to 2022-05-28 |
| Cases | cases Number of cases on that day for that state |
| Federal Information Processing | Federal Information Processing Standard State Code |
| Hospitalizations Cumulative | Cumulative Number of Hospitalizations per State (Used to calculate the hospitalization rate) |
| Deaths | deaths Number of deaths on that day for that state |
| Catchment | Catchment Different states in US |
| Age-Category | Age-Category Age Categories ranging from 0-85+ years. |
| Weekly-Rate | Weekly Rate Weekly Hospitalization Rate of different age groups |
| Medical Condition | Medical condition Different Medical conditions of Hospitalized adults |
| Hospitalized Adults | Total Number of Hospitalized Adults for different Medical Conditions |

Table 1: Attributes and Descriptions of all the datasets used for the experiments

## 2.2 DATA WRANGLING

Since the data may contain extraneous information, missing values, and inconsistent values, data wrangling is an important step. In order to be less error-prone and to avoid needless fallbacks caused by inconsistent data that are detected at subsequent processes, you need to have clean data. For our visualizations, we had to compute various parameters utilizing the columns that already existed in various datasets.

For instance, in order to assess the trend for each state, we needed to determine the adult hospitalisation rate. However, the hospitalization cumulative and positive cases columns were included in the dataset. For a better display, we thus had to compute and add the hospitalization rate column for each datapoint.



| | date | state | fips | cases | deaths | fatality_rate% |
|---|---|---|---|---|---|---|
| 55833 | 2022-12-04 | Virginia | 51 | 2153223 | 22582 | 1.048753 |
| 55834 | 2022-12-04 | Washington | 53 | 1859858 | 14739 | 0.792480 |
| 55835 | 2022-12-04 | West Virginia | 54 | 615332 | 7740 | 1.257858 |
| 55836 | 2022-12-04 | Wisconsin | 55 | 1929331 | 15684 | 0.812924 |
| 55837 | 2022-12-04 | Wyoming | 56 | 180925 | 1938 | 1.071162 |

Fig 2: Calculating the Fatality Rate Column

## 2.3 VISUALIZATIONS

The many trends in the data must be understood and analyzed, and data visualization is essential for this. We attempt to study the many patterns and causes contributing to the rise in the number of covid cases using line graphs, bar graphs, and tree maps. The visualizations were made using Tableau and Python modules like matplotlib, seaborn, and plotly.

We have created different graphs which have been explained below:

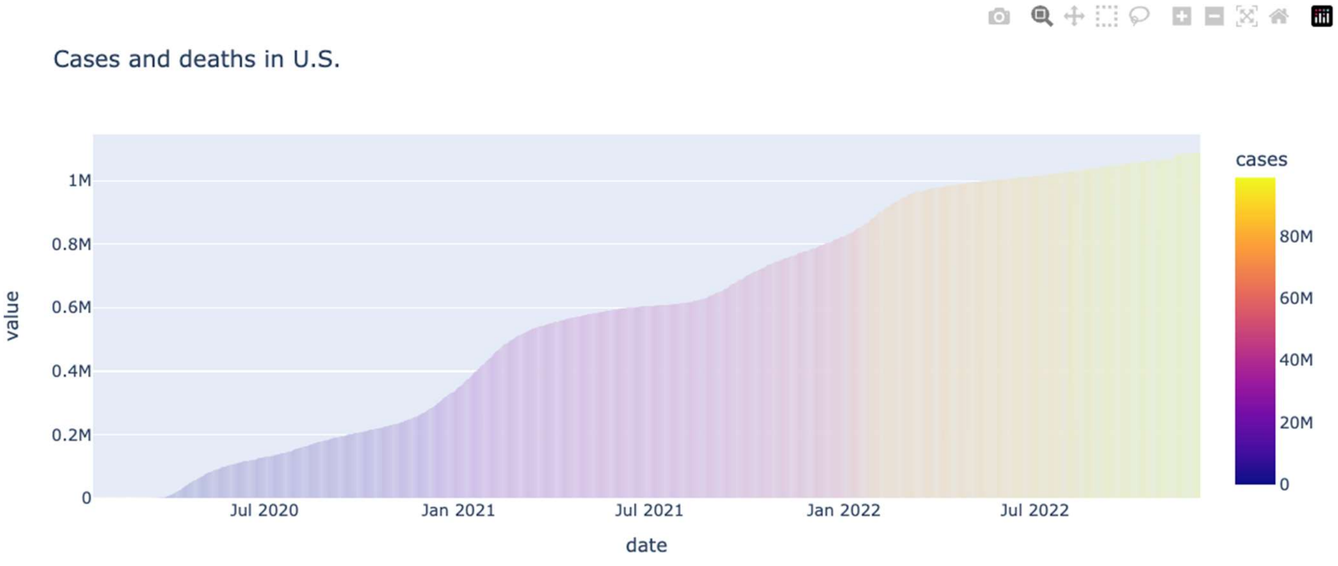
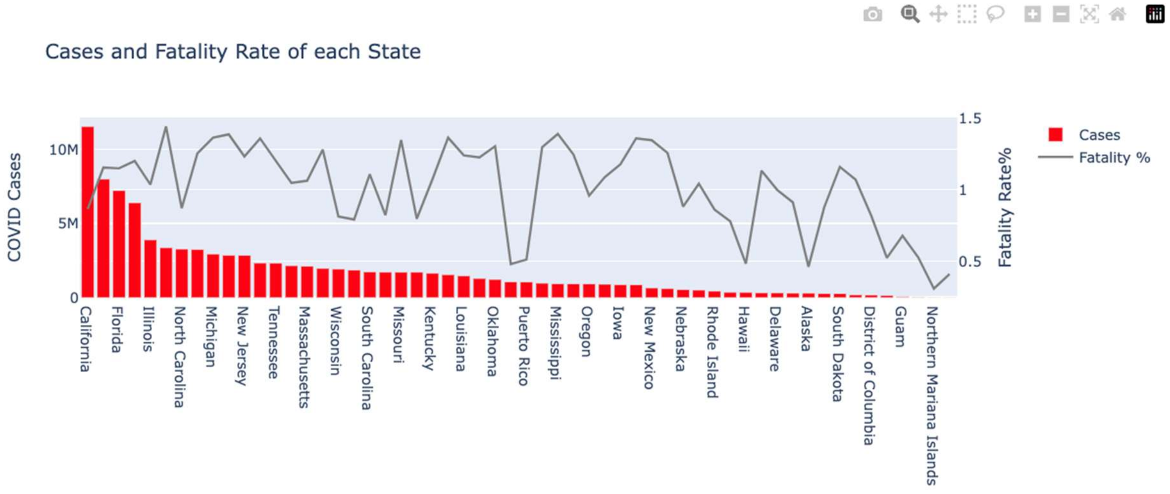


Fig 3: Cases and Deaths wrt Date

The number of cases and fatalities in the US from 2020 to 2022 is depicted in Figure 3. The first part of the graph, from October 2020 to February 2021, is what we'll examine. We can see that the number of instances increased significantly over that time. The fact that these are winter months with cold, dry weather should be noted. The rise in instances is caused by cold and dry weather, which makes it easier for the virus to live longer, alters the size and make-up of virus-carrying droplets, and weakens the defences of your respiratory tract. Additionally, the majority of activities migrate indoors during the winter, making it more difficult to maintain a safe social distance and a healthy airflow.



F

Fig 4: Cases and Fatality Rate of Each State

As shown in figure 4 above, New York has a relatively high mortality rate when compared to other large states, whereas Utah has a low death rate. This could be because New York City has the highest population density of any US metro region with over 100,000 residents. Different regions' mortality rates have been significantly impacted by location and population density.

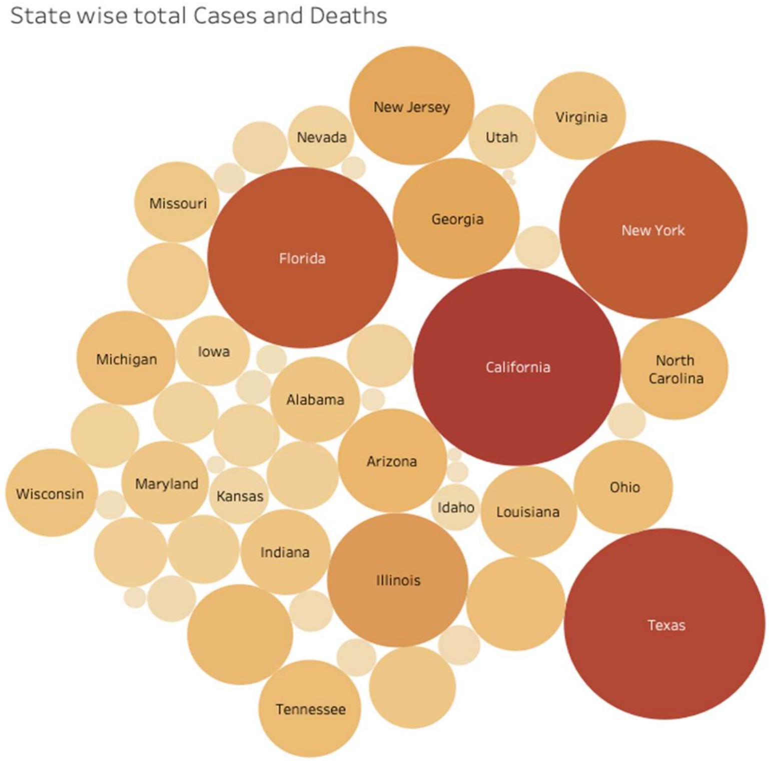


Fig 5: Total Cases for each state

The total number of cases for each US state are displayed in Figure 5. We can see that the most cases were found in California, Texas, Florida, and New York.

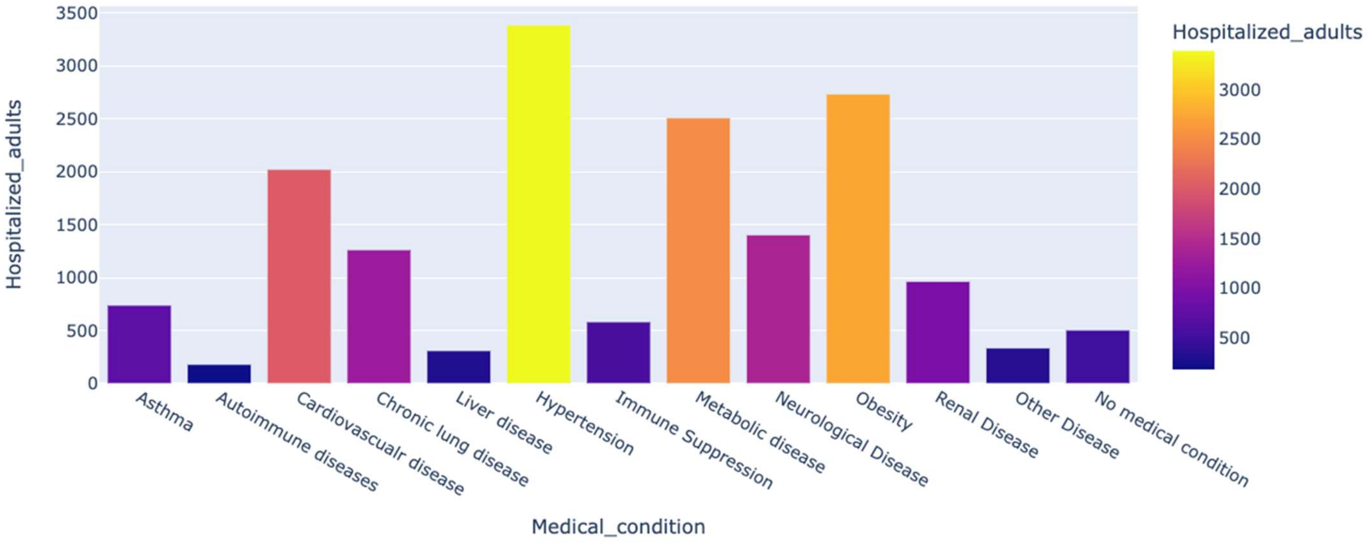


Fig 6: Classification of hospitalized adults wrt Medical conditions

Adult patients in hospitals are related to underlying medical issues, as seen in Figure 6. We comprehend the connection between hospitalised patients' underlying illnesses and those circumstances. One of the most common medical disorders among hospitalised patients was hypertension and obesity. The second most important illnesses were those related to the metabolism and the cardiovascular system. Figures 7 and 6 together support the idea that covid 19 was more detrimental to the elderly and people with underlying medical issues.

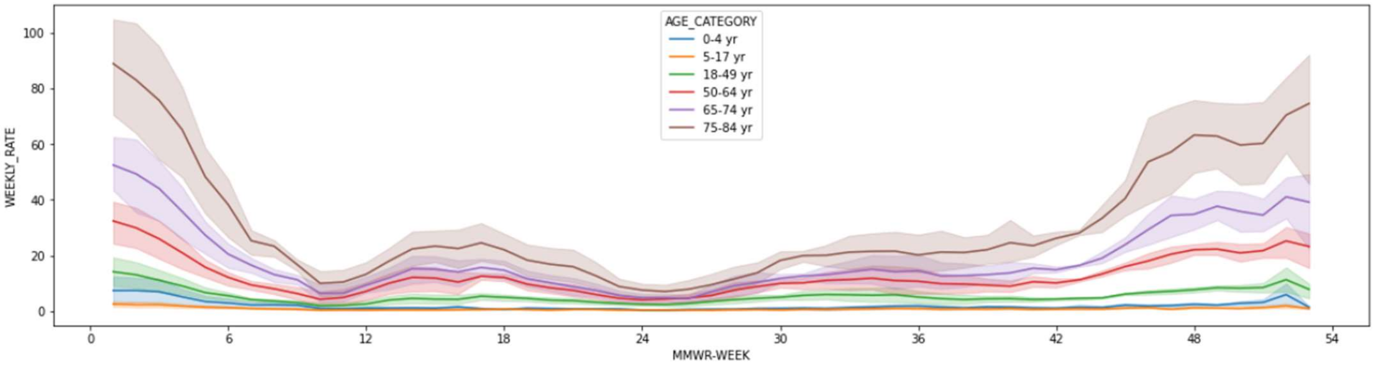


Fig 7: Weekly hospitalization rate for different age categories

Figure 7 displays the Covid 19 patients' weekly hospitalization rates in the US for various age groups. This graph demonstrates how the elderly population over 65 needed much more hospital treatment as a result of COVID 19. This may be because they have a high likelihood of having a condition that renders them vulnerable to the COVID virus and need medical attention.

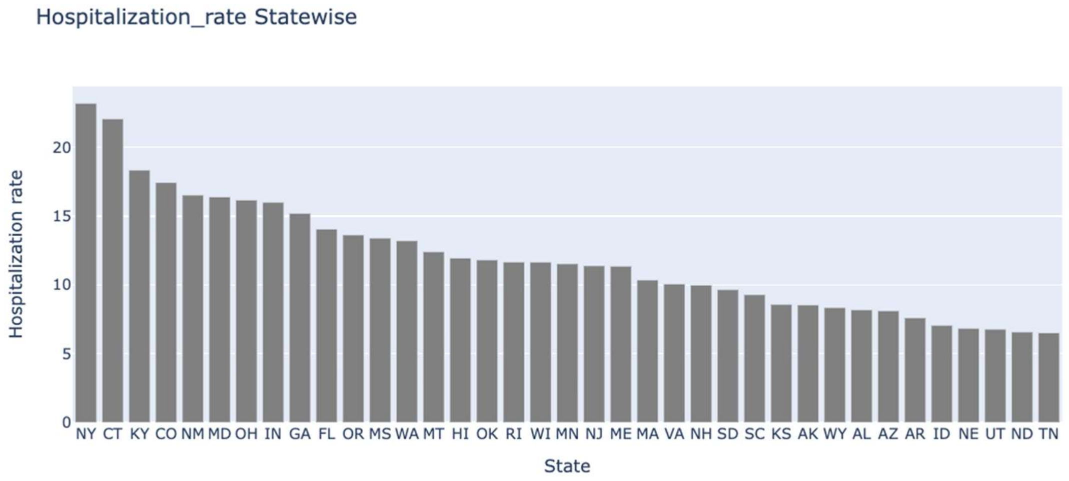


Fig 8: Hospitalization rate Statewise

The hospitalisation rate for each state is depicted in Figure 8. There may be a variety of causes for a patient to be hospitalised as a result of Covid. The patient may have underlying illnesses that worsen Covid's effects on the body. The patient's location and the area's population density are two influences. The graph above shows that New York has an extremely high hospitalisation rate.

## 2.4 TIME SERIES MODELS

A time series is a group of data points that have been catalogued chronologically in mathematics. Sequences of discrete-time data make up time series data. Time is frequently the independent variable in time series, and a forecast of the future is frequently the objective.

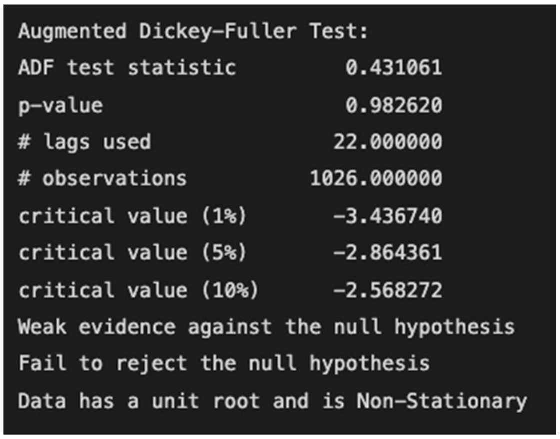
In an effort to forecast the number of COVID cases in the United States in the future, we performed time series analysis. We have forecasted values up to 100 days in the future for the dates 2020-01-21 to 2022-05-01 in our dataset.

Autoregressive Integrated Moving Average (ARIMA) and Seasonal Autoregressive Integrated Moving Average are the two models on which we have focused (SARIMA). There are measures we must take before employing the models. Finally, we used the LSTM model. However, the error rate was very high so we chose to use this to make long term predictions as the prediction line was very close to the base line.

**Data preprocessing:** The initial step It's crucial to constantly make sure that dates are utilized as index values and are recognized by Python as a real "date" object when working with time series data in Python. Using the to datetime function or the pandas date stamp, we can do this.

**Check for stationarity:** Second, most time series models demand steady data. If the mean, variance, and covariance of a time series do not change during the course of the series, the time series is said to be stationary. Plotting the data, doing a visual analysis, and employing a statistical test are the official techniques to verify this.

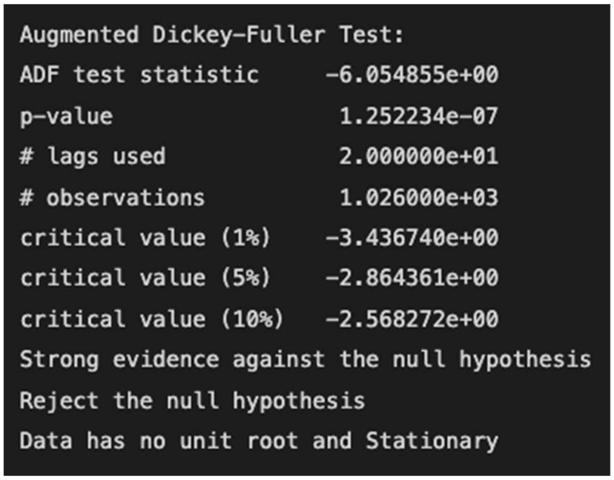
We must verify for stationarity after finishing the first step of transforming the date column to a date time object. For this check, we employ the Augmented Dickey Fuller Test (ADF Test). An often-used statistical test to determine whether a particular Time series is stationary or not is the Augmented Dickey Fuller test (ADF Test). When examining the stationary of a series, it is one of the statistical tests that is most frequently applied. To run the ADF test and return the ADF report, we built a function. We deduced that the data is not steady from the report.



```
Augmented Dickey-Fuller Test:
ADF test statistic        0.431061
p-value                   0.982620
# lags used              22.000000
# observations         1026.000000
critical value (1%)      -3.436740
critical value (5%)      -2.864361
critical value (10%)     -2.568272
Weak evidence against the null hypothesis
Fail to reject the null hypothesis
Data has a unit root and is Non-Stationary
```

Fig 9: First ADF report

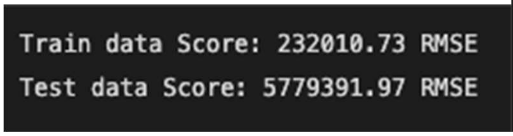
As a result, we obtain the ARIMA orders using Python's auto arima function. This procedure returns the ARIMA order that should be applied to our dataset. For our dataset, the order was as follows: (2,2,2). We now difference our data twice using the stats models library diff before rechecking stationarity.



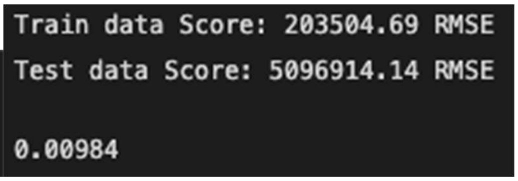
```
Augmented Dickey-Fuller Test:
ADF test statistic     -6.054855e+00
p-value                 1.252234e-07
# lags used             2.000000e+01
# observations          1.026000e+03
critical value (1%)    -3.436740e+00
critical value (5%)    -2.864361e+00
critical value (10%)   -2.568272e+00
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and Stationary
```

Fig 10: Second ADF report

Finally, we trained for the LSTM model and as we can see, that it had a very high error rate so we moved on to the next step of plotting graphs. Having said that, this model can be used to make long term predictions as the difference between the prediction and the base line is very little.



```
Train data Score: 232010.73 RMSE
Test data Score: 5779391.97 RMSE
```



```
Train data Score: 203504.69 RMSE
Test data Score: 5096914.14 RMSE

0.00984
```

Fig 10.1: LSTM Report

Fig 10.2: LSTM windows report

The next stage in fitting an ARIMA model is to establish if AR or MA terms are required to correct any autocorrelation that persists in the differenced series after a time series has been stationarized by differencing. You can hazard a guess as to how many AR and/or MA terms are required by examining the differenced series' autocorrelation function (ACF) and partial autocorrelation (PACF) graphs.

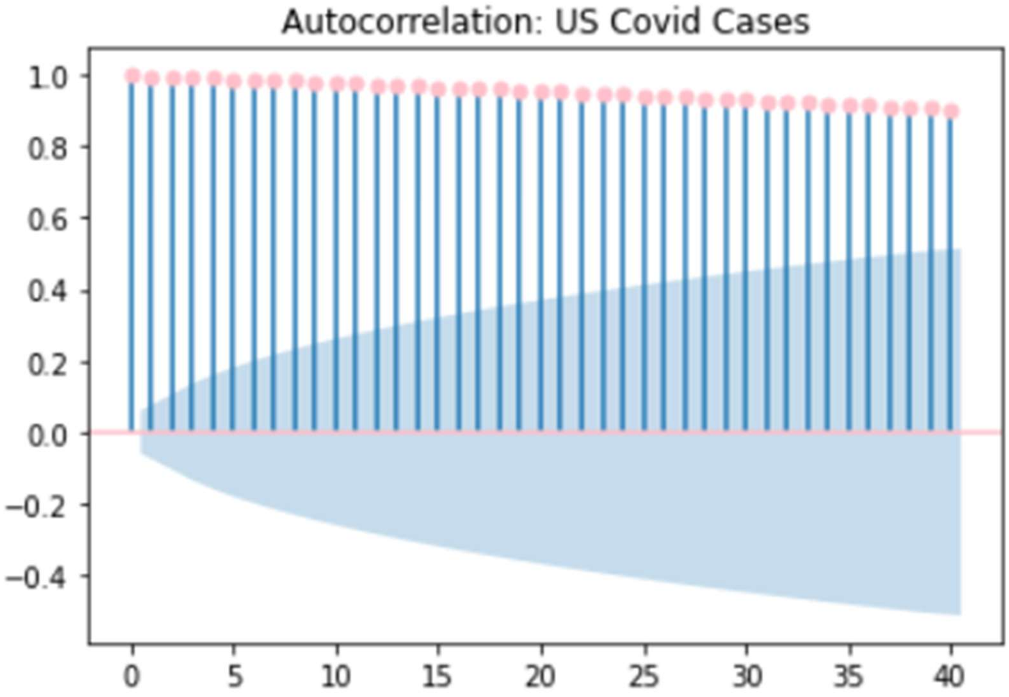
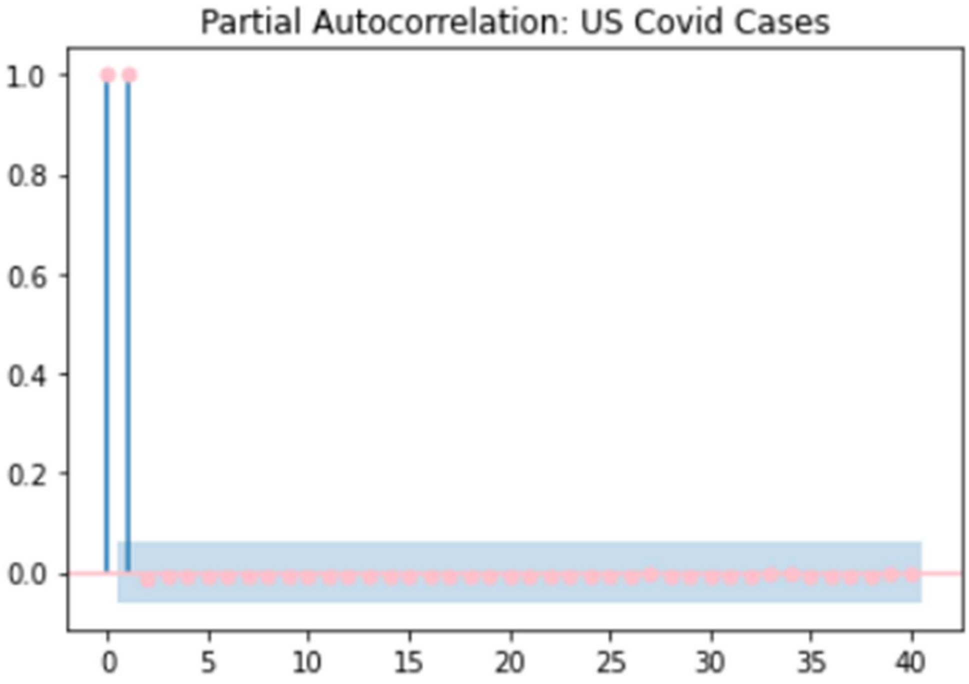


Fig 12: Autocorrelation Plot



` Fig 13: Partial Autocorrelation Plot

The two graphs show that the AR component should be more significant than the MA component. We have now created an ARIMA model for our dataset. A type of statistical models known as ARIMA models are used to analyse and predict time series data, where AR stands for Autoregression, I for Integration, and MA for Moving Average. The following is a definition of the ARIMA model's parameters:

- p: The lag order, commonly known as the number of lag observations contained in the model.
- d: The degree of differencing, also known as the number of times the raw observations are differed.
- q: The moving average window size, commonly known as the moving average order.

Additionally, we sought to determine whether the frequency of COVID cases is seasonal. As a result, we also used the SARIMA model, which allows us to include a seasonal component.

## 3 Results

We forecasted the values for the next 100 days using the ARIMA, SARIMA  and LSTM models. The model summaries for both models are shown in the following figures.

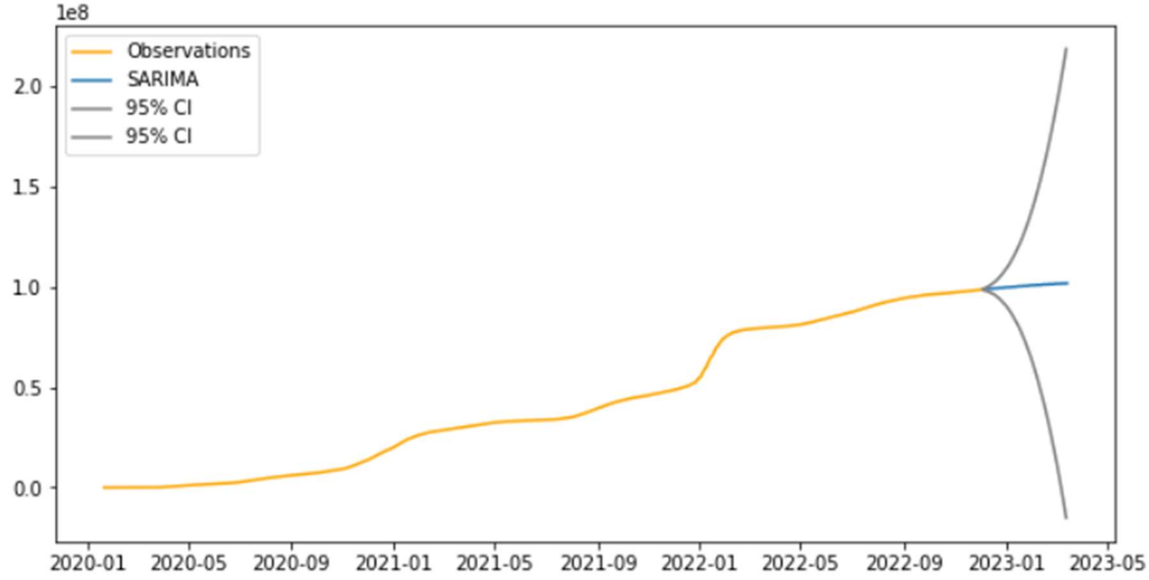


Fig 14: Forecast of SARIMA

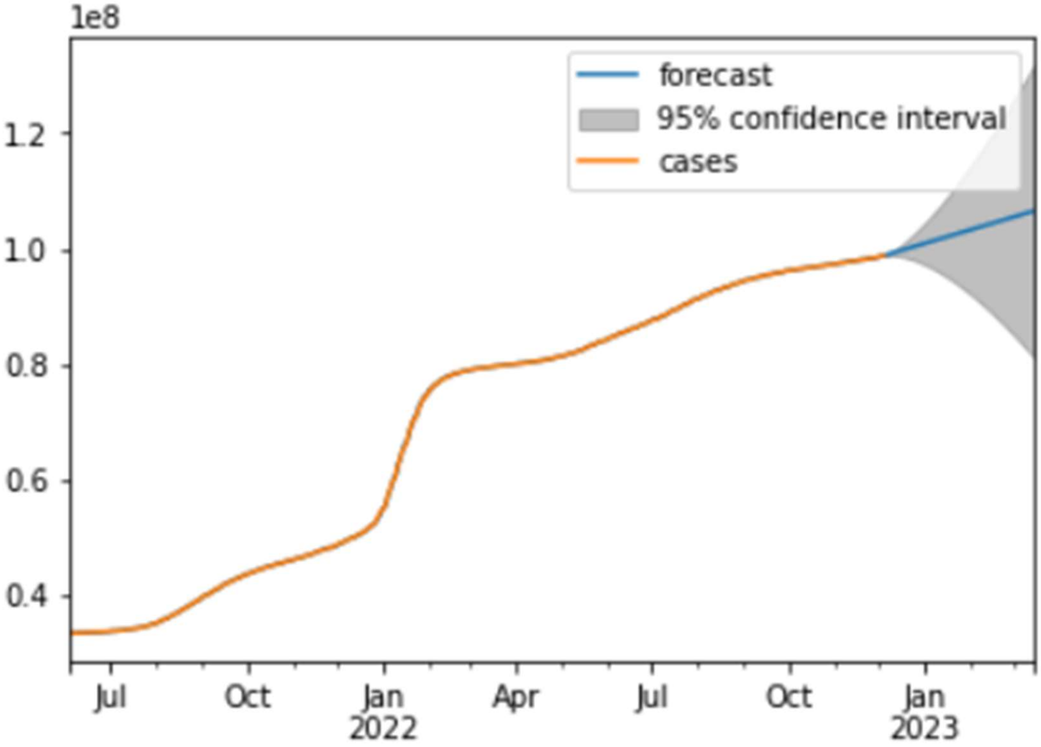


Fig 15: Forecast of ARIMA

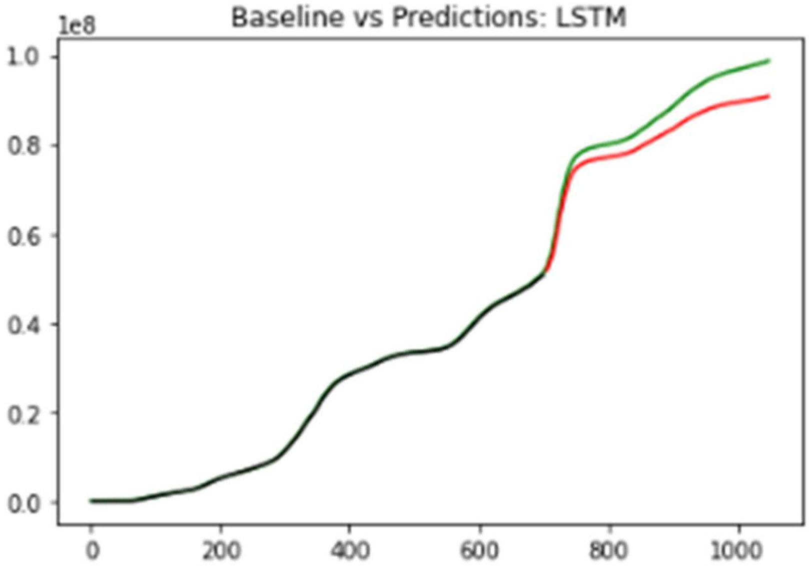


Fig 16. Forecast of LSTM

The error rate for the ARIMA and SARIMA models was 0.00098. We trained for the LSTM model and as we can see, that it had a very high error rate so we moved on to the next step of plotting graphs. Having said that, this model can be used to make long term predictions as the difference between the prediction and the base line is very little. However, as you can see in the graph, the confidence interval gets bigger with time. The influence of seasonality was not evident in the forecasts for SARIMA since the number of instances frequently relies on the country's current policy, such as a nationwide lockdown. In the future, we may validate the feasibility of SARIMA using a monthly dataset.

## 4 Discussion and Conclusion

We may infer from the findings that these models are appropriate for short-term forecasts of COVID instances. We underestimate how intricate pandemics really are. The accuracy of the model is impacted by a variety of other variables, including viral mutations, vaccine production efficiency, population vaccination rates, etc. The visualizations in this report helped us to understand that the impacts of the virus varied depending on the state and the population, as well as from the visualizations themselves. Governments should have plans in place to assist these regions if a similar epidemic strikes in the future, well before things get out of hand.

## 5 References

[1] Andrea Remuzzi, Prof, EngD, Giuseppe Remuzzi, Prof, MD COVID-19 and Italy: what next? - PMC, 2020 11-17 April

[2] Rob J Hyndman and George Athanasopoulos, "8.9 Seasonal ARIMA models", Forecasting: principles and practice, May 2015.

[3] Navid Mashinchi, Predicting number of Covid19 deaths using Time Series Analysis (ARIMA MODEL) | by Navid Mashinchi | Towards Data Science, Sep 22, 2020

[4] Qiang Wang, Shuyu Li and Rongrong Li, "Forecasting Energy Demand in China and India: Using Single-linear Hybrid-linear and Non-linear Time Series Forecast Techniques", Energy Elsevier, vol. 161, no. C, pp. 821-831, 2018.

[5] G E P Box and G M Jenkins, "Time Series Analysis: Forecasting and Control [J]", Journal of the American Statistical Association, vol. 68, no. 342, pp. 199-201, 1970.

[6] David W. S. Wong, Yun Li, Spreading of COVID-19: Density matters, Dec 23 2020