**COVID-19 Analysis in the United States**

Pranav Addipalli

Vinit Horekari

Kirat Saran

Syracuse University

Syracuse, NY, 13210 USA

**Statement of Purpose:**

COVID – 19 outbreak started in Wuhan, China in December 2019 when 50 people developed pneumonia-like symptoms. Several cases of COVID infection were reported after that in that district and a lockdown was imposed. No one knew what it was exactly and how it spread like if it was airborne or spread through contact. There was a huge spike in cases in the area which lead to it being declared as an epidemic by World Health Organization (WHO). Cases suddenly spread throughout the globe because of travel and eventually it was declared as a pandemic by WHO. Scientists and doctors in this phase were still trying to figure out its origin and cure along with its long term effects. Developing a vaccine to cure it was soon being discussed.

In our project, we aim to track and analyse the outbreak in the United States of America by focusing on total RTPCR tests conducted, number of hospitalizations that were led by the outbreak, number of people who recovered from COVID, the mortalities because of the outbreak, etc.

**Analysis and Predictions:**

We plan on exploring the cases reported to understand how rapidly the virus spread exponentially throughout the United States and its impacts. This when corelated to other variables will give us a fair understanding on how the outbreak turned into a life-changing pandemic.

Finding out the trends between deaths and cases reported helps us to understand this.

**Data Used:**

We have used the data from an official source, “covidtracking.com” which gives a day-wise information about deaths, total cases recorded, positive count, recovered patients, patients in ICU and patients on ventilators, etc. Our dataset spans from 13th Jan 2020 just around the time when first covid case came to the country to 7th March 2021 that is just before the end of the first wave of covid.

The main variables that we have focused on to analyse and predict the outcomes are:

* death: Total fatalities with confirmed OR probable COVID-19 case diagnosis
* totalTestResults: At the national level, this metric is a summary statistic which, because of the variation in test reporting methods, is at best an estimate of US viral (PCR) testing.
* recovered: Total number of people that are identified as recovered from COVID-19. States provide very disparate definitions on what constitutes a “recovered” COVID-19 case.

**Data Preparation and Pre-Processing:**

To prepare the data, we need to clean the data retrieved from covidtracking.com. The cleaning is done by checking

* the data types (Fig. 1)
* null values (Fig. 2)
* skewness
* statistical importance
* outliers

This helps us understand which variables need to be manipulated and imputed. We also need to explore all the columns to understand relation between the attributes and get a deeper understanding of the dataset.

The data cleaning is done by replacing the null values with zeroes since the data missing is usually zero. Next, we checked for duplicate rows and values. We discovered that there were no duplicate values in our dataset. The next step we did was that we converted date column’s type to datetime from object to set it as our dataset’s index() because it makes it easier to do timeseries analysis. This is done using the NumPy library imported at the beginning of processing. Then, the count of rows and columns is taken after each step to check how much data has been expunged.

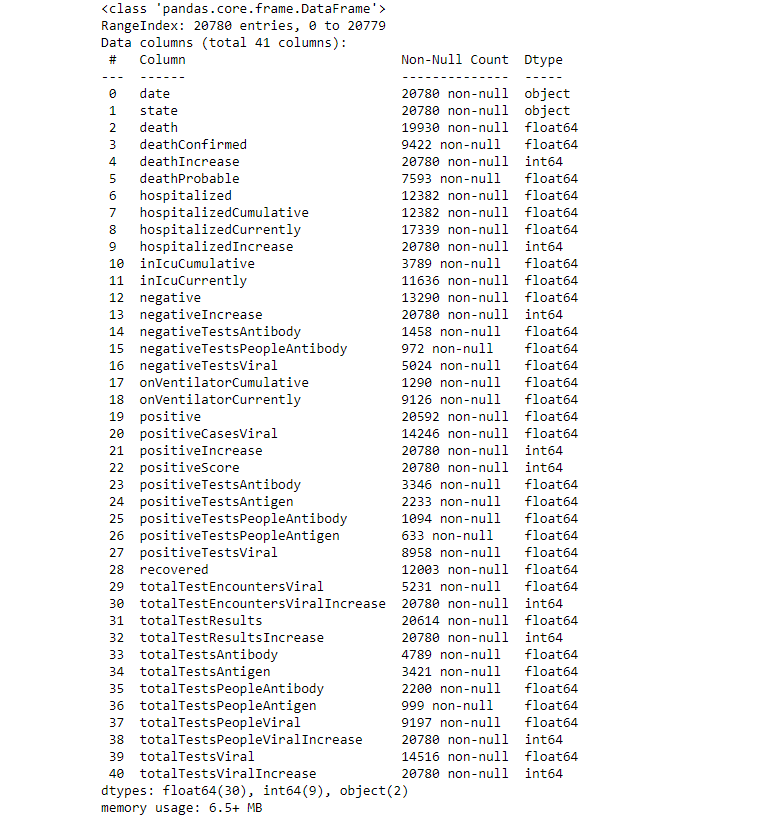


Fig. 1

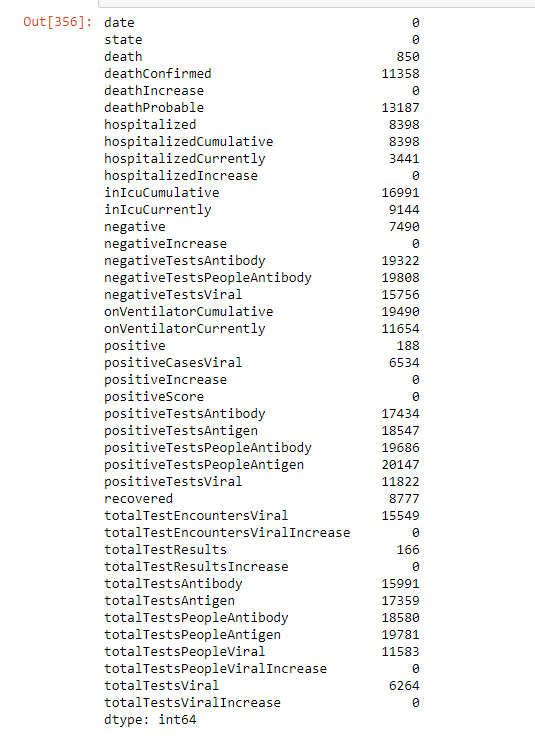
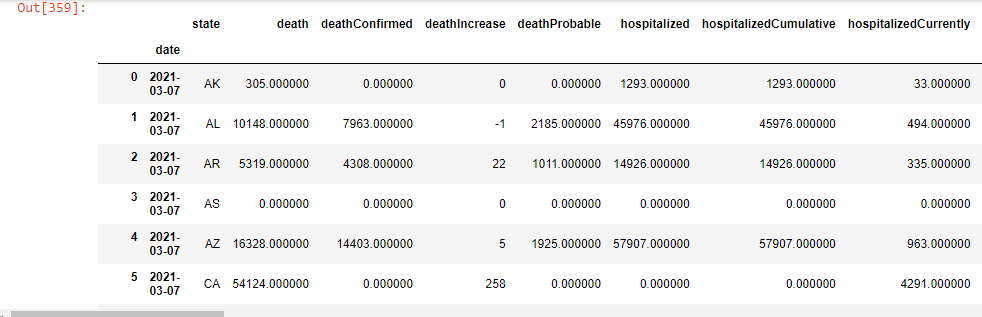


Fig. 2

Fig. 3

**Visualization and Exploratory Analysis**

To understand the data and establish a relationship between the attributes and the dataset we must visualize the same. These attributes help define the depth of understanding of data through visualization. Data Visualization helps us understand the data better because:

* Python provides myriad resources to better plot and visualize the data according to our needs.
* It has tools built in statistical functions, which reveal hidden patterns in the data set.
* It has functions to visualize matrices of data, which become very important when visualizing large data sets

Visualization is done to further understand the hidden patterns in the dataset, which helps in future processing and modelling. We have used the following visualizations:

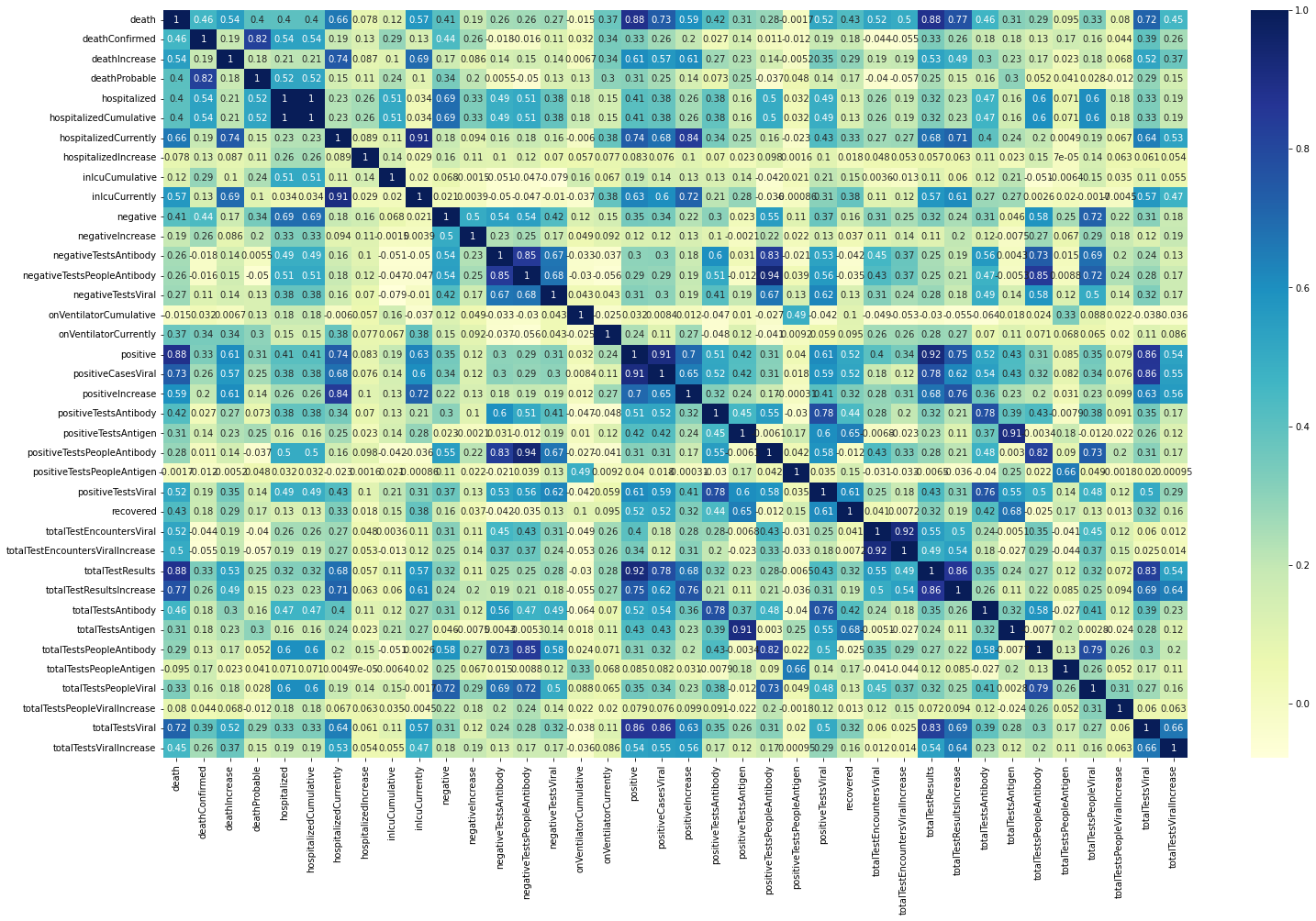


Fig. 4

The heat map in Figure 4 helps us understand the correlation between the different attributes in the dataset. This can be used later for machine learning models and statistical modelling.

Chart, box and whisker chart

Description automatically generated

Fig. 5

Chart, box and whisker chart

Description automatically generated

Fig. 6

Chart, box and whisker chart

Description automatically generated

Fig. 7

Chart, box and whisker chart

Description automatically generated

Fig. 8

Chart, box and whisker chart

Description automatically generated with medium confidence

Fig. 9

Chart, box and whisker chart

Description automatically generated

Fig. 10

Chart

Description automatically generated

Fig. 11

The box plots from Figure 6 to Figure 11 illustrate that the data contain outliers that must be removed in order to describe the data accurately. The interquartile function can be used to remove these outliers from the data. This interquartile function necessitates first dividing the data into quartiles and then removing outliers with the provided function. This aids in the reduction of data processing errors.

A picture containing text, crossword puzzle

Description automatically generated

Fig. 12

The figure 12 is histograms of all the columns, it can be observed that most of the columns are positively skewed, here the mean of positively skewed data will be greater than the median.

Chart

Description automatically generated

Fig. 13

The plot in Figure 13 shows the number of hospitalizations to the number of deaths due to the virus. This helps understand the mortality rate of the population due to COVID-19 virus.

Even though the hospitalisations seems to be decreasing over time, the deaths are constant for a long period of time.

Chart, histogram

Description automatically generated

Fig. 14

Figure 14 is a plot representing the distribution of number of cases over the time, the Y axis has number of positive cases and date on the X axis

A picture containing graphical user interface

Description automatically generated

Fig. 15

Figure 15 is a plot representing the distribution of number of cases over the time, the Y axis has number of recovered cases and date on the X axis.

This plot depicts that the recovered cases steeply started rising from January 2021 onwards. The reason might be due the vaccination drive.

Graphical user interface, line chart

Description automatically generated

Fig. 16

Figure 16 is a line plot representing the weekly number of cases, the Y axis has number of cases, and week number on the X axis.

Through week 25, the confirmed cases started rising exponentially while the recovery rate is increasing gradually. This might be due to the fact that many people had already contracted the virus and along with that due to that vaccination rollout, people had already started developing antibodies.

Chart, line chart

Description automatically generated

Fig. 17

Figure 17 is a line plot representing the weekly number of cases, the Y axis has number of cases, and date on the X axis.

Confirmed cases started increasing with a huge spike starting from November 2020. This might be due to the fact that holiday season begins around the same time.

Chart, line chart

Description automatically generated

Fig. 18

Figure 18 is a line plot representing the recovery rate over time, the Y axis has Recovery Rate, and date on the X axis.

Recovery Rate= (Number of Recoverd Cases / Number of Confirmed Cases) x 100.

Chart

Description automatically generated

Fig. 19

Figure 19 is a line plot representing the mortality rate over time, the Y axis has mortality Rate, and date on the X axis.

Mortality rate = (Number of Death Cases / Number of Confirmed Cases) x 100.

Chart, histogram

Description automatically generated

Fig.20

Figure 20 shows the daily increase in the different types of cases over the time. Y axis has number of cases, and date on the X axis.

Chart, histogram

Description automatically generated

Fig. 21

Figure 21 shows the 7 days rolling mean for the different types of cases over the time. The Y axis has number of cases, and date on the X axis.

A picture containing table

Description automatically generated

Fig. 22

Figure 22 shows the date wise growth factor for the different types of cases over the time. The Y axis has the growth factor of cases, and date on the X axis.

Graphical user interface

Description automatically generated with medium confidence

Fig. 23

Figure 23 shows the date wise growth factor for the different types of cases over the time. The Y axis has the growth factor of cases, and date on the X axis.

Understanding for fig 18-fig23:

* Growth factor is the factor by which a quantity multiplies itself over time.
* Formula: Every day's new (Confirmed,Recovered,Deaths) / new (Confirmed,Recovered,Deaths) on the previous day.
* A growth factor above 1 indicates an increase correspoding cases.
* A growth factor above 1 but trending downward is a positive sign, whereas a growth factor constantly above 1 is the sign of exponential growth.
* A growth factor constant at 1 indicates there is no change in any kind of cases.

Graphical user interface

Description automatically generated with medium confidence Fig. 24

Figure 24 shows the number of days it took for the cases to double, high number of days is a positive sign.

* The doubling rate represents the number of days it takes for the number of COVID-19 cases to double, an indicator of how quickly cases are increasing.
* The doubling rate can also be applied to assess the trajectory of hospitalizations and deaths, providing key information on whether a region is slowing the spread of the disease.
* The longer the doubling rate, the slower the disease spreads and the flatter the curve becomes.
* Conversely, a shorter doubling rate indicates the disease is spreading more quickly.

Chart

Description automatically generated

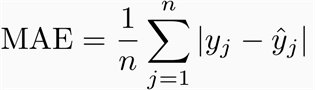
Fig. 25

Linear regression is a data plot that graphs the linear relationship between an independent and a dependent variable, here Confirmed cases i.e. positive is the dependent variable.

Model Evaluation Metrics is Mean Absolute Error(MAE).

For positive cases prediction MAE for the Linear Regression model is 2843 and for SVM it is 3593

**Mean Absolute Error (MAE):** MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It’s the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.



Graphical user interface

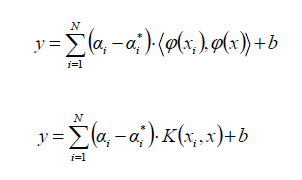
Description automatically generated

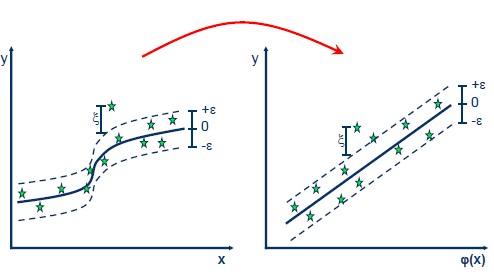
Fig. 26

Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. First of all, because output is a real number it becomes very difficult to predict the information at hand, which has infinite possibilities. In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem.

**Non-Linear SVR**

The kernel functions transform the data into a higher dimensional feature space to make it possible to perform the linear separation





Since the positive cases distribution is not linear, SVM can fit a curve unlike regression

Graphical user interface

Description automatically generated with low confidence

Fig. 27

Graphical user interface

Description automatically generated with medium confidence

Fig. 28

Chart

Description automatically generated

Fig. 29

Graphical user interface, application

Description automatically generated

Fig. 30

Results and Insights:

* The Healthy Recovery Rate implies the disease is curable.
* The only matter of concern is the exponential growth rate and spread of infection.
* The growth of Confirmed and Death Cases seems to have slowed down since the past few days. Which is a good sign.
* Rolling mean of confirmed cases is decreasing after a peak in Jan 2021.
* Doubling rate of positive cases is increasing which is also a good sign.
* Although the maximum number of hospitalisations are in Florida, more mortalities have occurred in the state of New York.

**References:**

* Lecture slides and ipynb files.
* Textbook
* Online resources websites:

1. Geeksforgeeks.org
2. stackoverflow.com
3. pandas.pydata.org
4. scikit-learn.org