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:

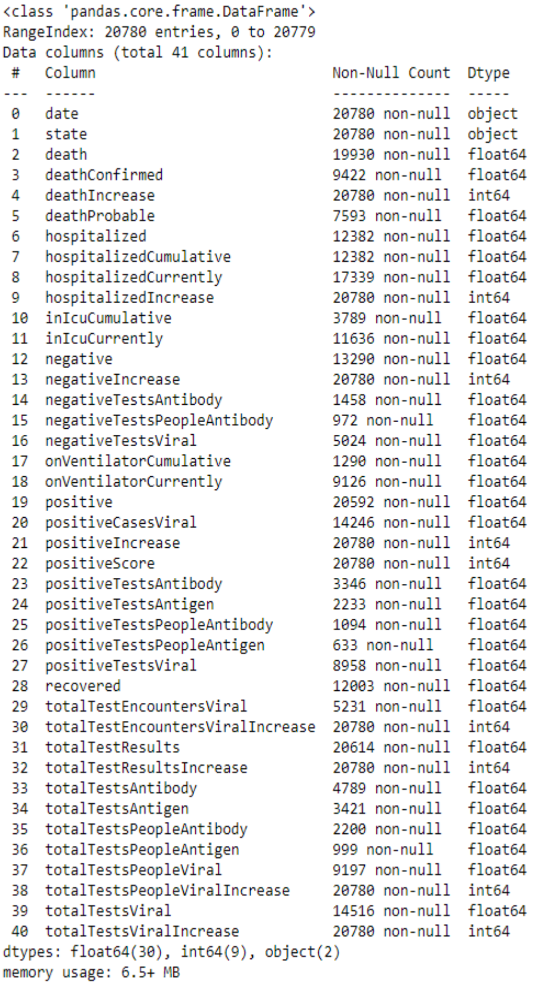
To prepare the data, we need to clean the data retrieved from covidtracking.com. The cleaning is done by checking

- the data types (Fig. 1)
- null values (Fig. 2)

- skewness
- statistical importance
- outliers

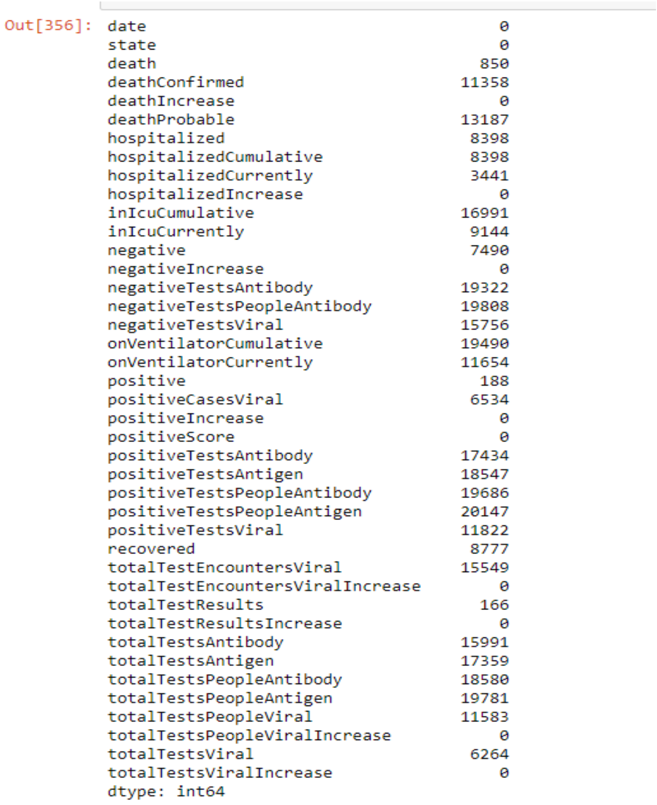
This helps us understand which variables need to be manipulated and imputed. We also need to explore all the columns to understand relation between the attributes and get a deeper understanding of the dataset.

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.



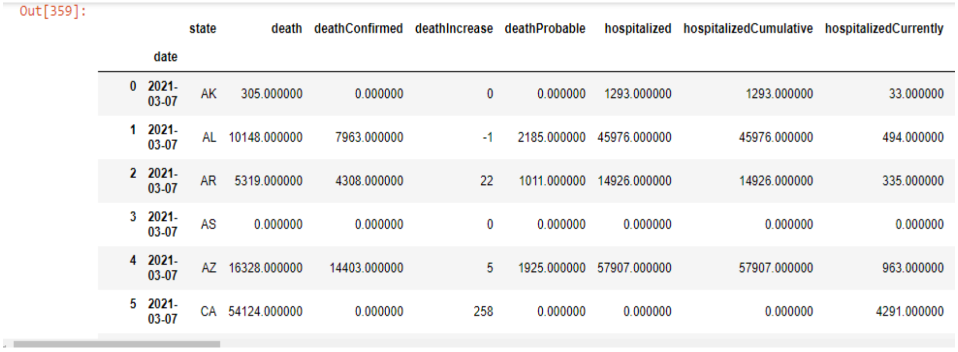
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20780 entries, 0 to 20779
Data columns (total 41 columns):
 #   Column                           Non-Null Count  Dtype  
---  ------                           --------------  -----  
 0   date                             20780 non-null  object 
 1   state                            20780 non-null  object 
 2   death                            19930 non-null  float64
 3   deathConfirmed                   9422 non-null   float64
 4   deathIncrease                    20780 non-null  int64  
 5   deathProbable                    7593 non-null   float64
 6   hospitalized                     12382 non-null  float64
 7   hospitalizedCumulative           12382 non-null  float64
 8   hospitalizedCurrently            17339 non-null  float64
 9   hospitalizedIncrease             20780 non-null  int64  
 10  inIcuCumulative                  3789 non-null   float64
 11  inIcuCurrently                   11636 non-null  float64
 12  negative                         13290 non-null  float64
 13  negativeIncrease                 20780 non-null  int64  
 14  negativeTestsAntibody            1458 non-null   float64
 15  negativeTestsPeopleAntibody      972 non-null    float64
 16  negativeTestsViral               5024 non-null   float64
 17  onVentilatorCumulative           1290 non-null   float64
 18  onVentilatorCurrently            9126 non-null   float64
 19  positive                         20592 non-null  float64
 20  positiveCasesViral               14246 non-null  float64
 21  positiveIncrease                 20780 non-null  int64  
 22  positiveScore                    20780 non-null  int64  
 23  positiveTestsAntibody            3346 non-null   float64
 24  positiveTestsAntigen             2233 non-null   float64
 25  positiveTestsPeopleAntibody      1094 non-null   float64
 26  positiveTestsPeopleAntigen       633 non-null    float64
 27  positiveTestsViral               8958 non-null   float64
 28  recovered                        12003 non-null  float64
 29  totalTestEncountersViral         5231 non-null   float64
 30  totalTestEncountersViralIncrease 20780 non-null  int64  
 31  totalTestResults                 20614 non-null  float64
 32  totalTestResultsIncrease         20780 non-null  int64  
 33  totalTestsAntibody               4789 non-null   float64
 34  totalTestsAntigen                3421 non-null   float64
 35  totalTestsPeopleAntibody         2200 non-null   float64
 36  totalTestsPeopleAntigen          999 non-null    float64
 37  totalTestsPeopleViral            9197 non-null   float64
 38  totalTestsPeopleViralIncrease    20780 non-null  int64  
 39  totalTestsViral                  14516 non-null  float64
 40  totalTestsViralIncrease          20780 non-null  int64  
dtypes: float64(30), int64(9), object(2)
memory usage: 6.5+ MB
```

Fig. 1



```
Out[356]: date                                   0
          state                                  0
          death                                850
          deathConfirmed                     11358
          deathIncrease                          0
          deathProbable                      13187
          hospitalized                        8398
          hospitalizedCumulative              8398
          hospitalizedCurrently               3441
          hospitalizedIncrease                   0
          inIcuCumulative                    16991
          inIcuCurrently                      9144
          negative                            7490
          negativeIncrease                       0
          negativeTestsAntibody              19322
          negativeTestsPeopleAntibody        19808
          negativeTestsViral                 15756
          onVentilatorCumulative             19490
          onVentilatorCurrently              11654
          positive                             188
          positiveCasesViral                  6534
          positiveIncrease                       0
          positiveScore                          0
          positiveTestsAntibody              17434
          positiveTestsAntigen               18547
          positiveTestsPeopleAntibody        19686
          positiveTestsPeopleAntigen         20147
          positiveTestsViral                 11822
          recovered                           8777
          totalTestEncountersViral           15549
          totalTestEncountersViralIncrease       0
          totalTestResults                     166
          totalTestResultsIncrease               0
          totalTestsAntibody                 15991
          totalTestsAntigen                  17359
          totalTestsPeopleAntibody           18580
          totalTestsPeopleAntigen            19781
          totalTestsPeopleViral              11583
          totalTestsPeopleViralIncrease          0
          totalTestsViral                     6264
          totalTestsViralIncrease                0
          dtype: int64
```

Fig. 2



Out[359]:

		state	death	deathConfirmed	deathIncrease	deathProbable	hospitalized	hospitalizedCumulative	hospitalizedCurrently
	date								
0	2021-03-07	AK	305.000000	0.000000	0	0.000000	1293.000000	1293.000000	33.000000
1	2021-03-07	AL	10148.000000	7963.000000	-1	2185.000000	45976.000000	45976.000000	494.000000
2	2021-03-07	AR	5319.000000	4308.000000	22	1011.000000	14926.000000	14926.000000	335.000000
3	2021-03-07	AS	0.000000	0.000000	0	0.000000	0.000000	0.000000	0.000000
4	2021-03-07	AZ	16328.000000	14403.000000	5	1925.000000	57907.000000	57907.000000	963.000000
5	2021-03-07	CA	54124.000000	0.000000	258	0.000000	0.000000	0.000000	4291.000000

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:

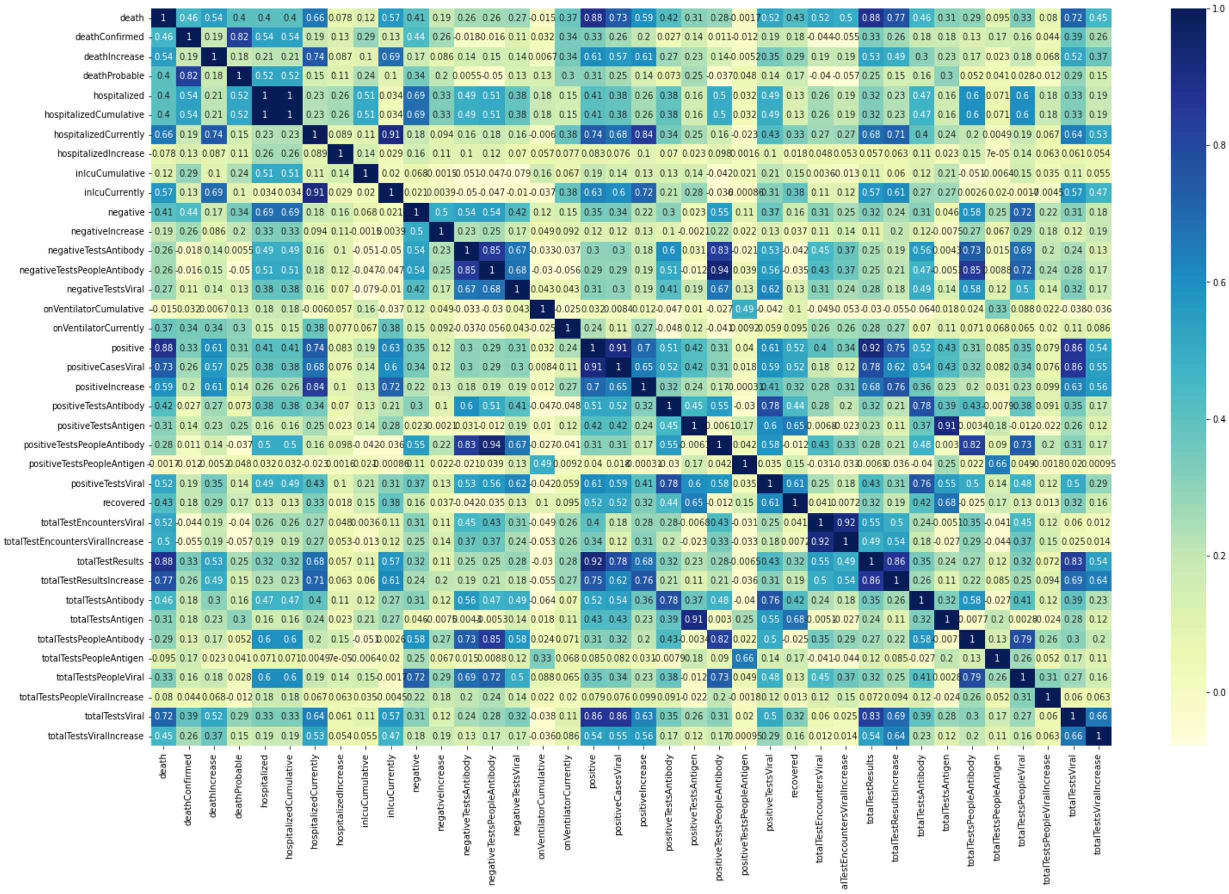


Fig. 4

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.

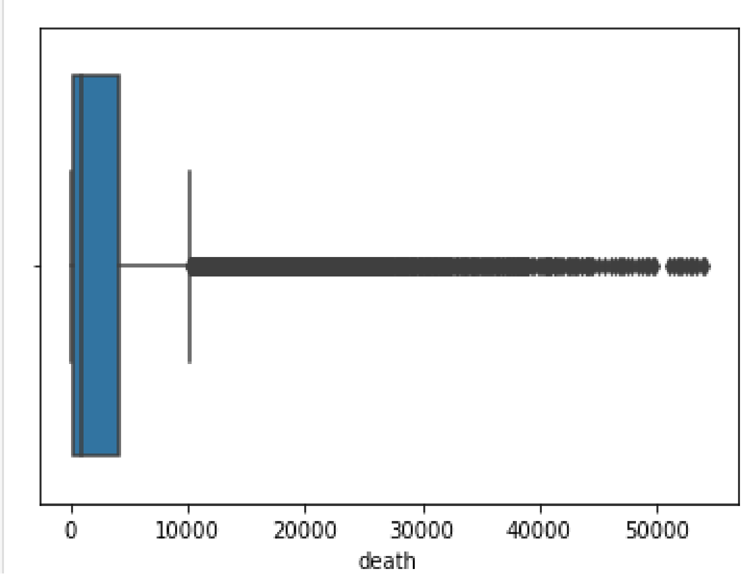


Fig. 5

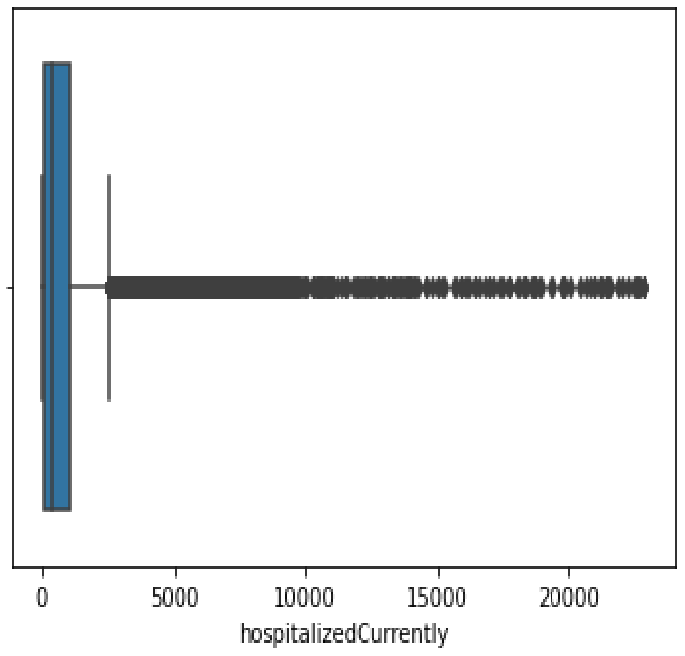

0
5000
10000
15000
20000
hospitalizedCurrently

Fig. 6

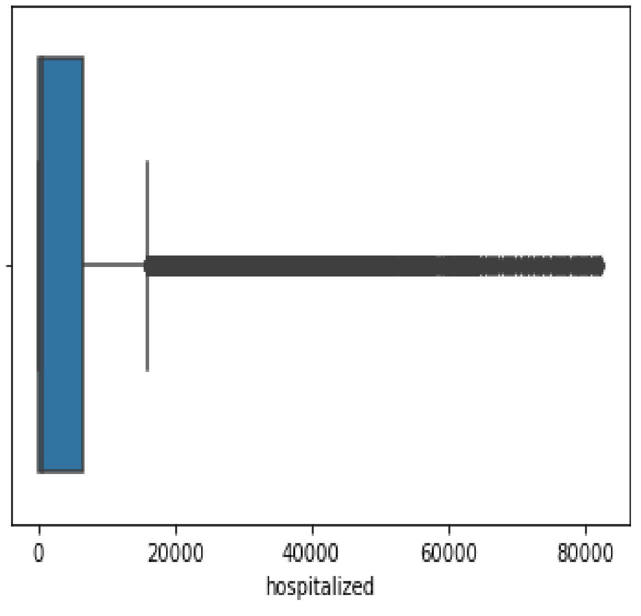

0
20000
40000
60000
80000
hospitalized

Fig. 7

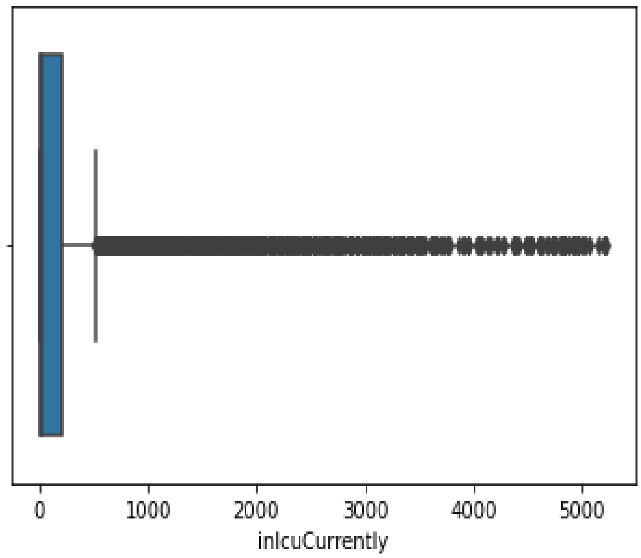

0
1000
2000
3000
4000
5000
inIcuCurrently

Fig. 8

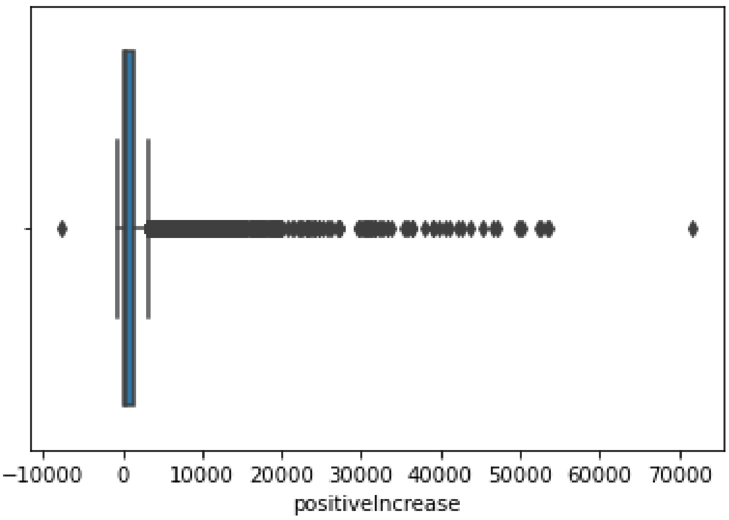

−10000
0
10000
20000
30000
40000
50000
60000
70000
positiveIncrease

Fig. 9

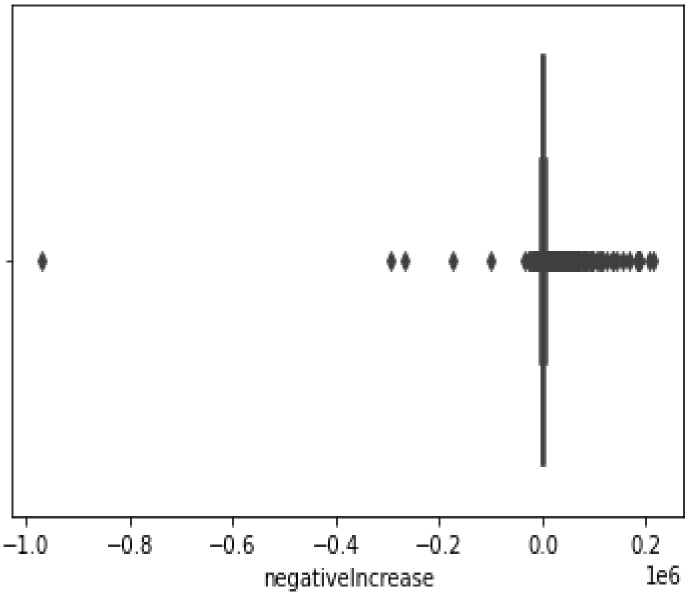


Fig. 10

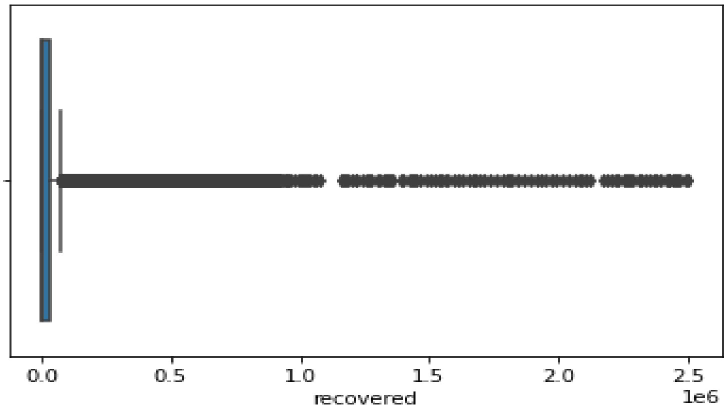


Fig. 11

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.

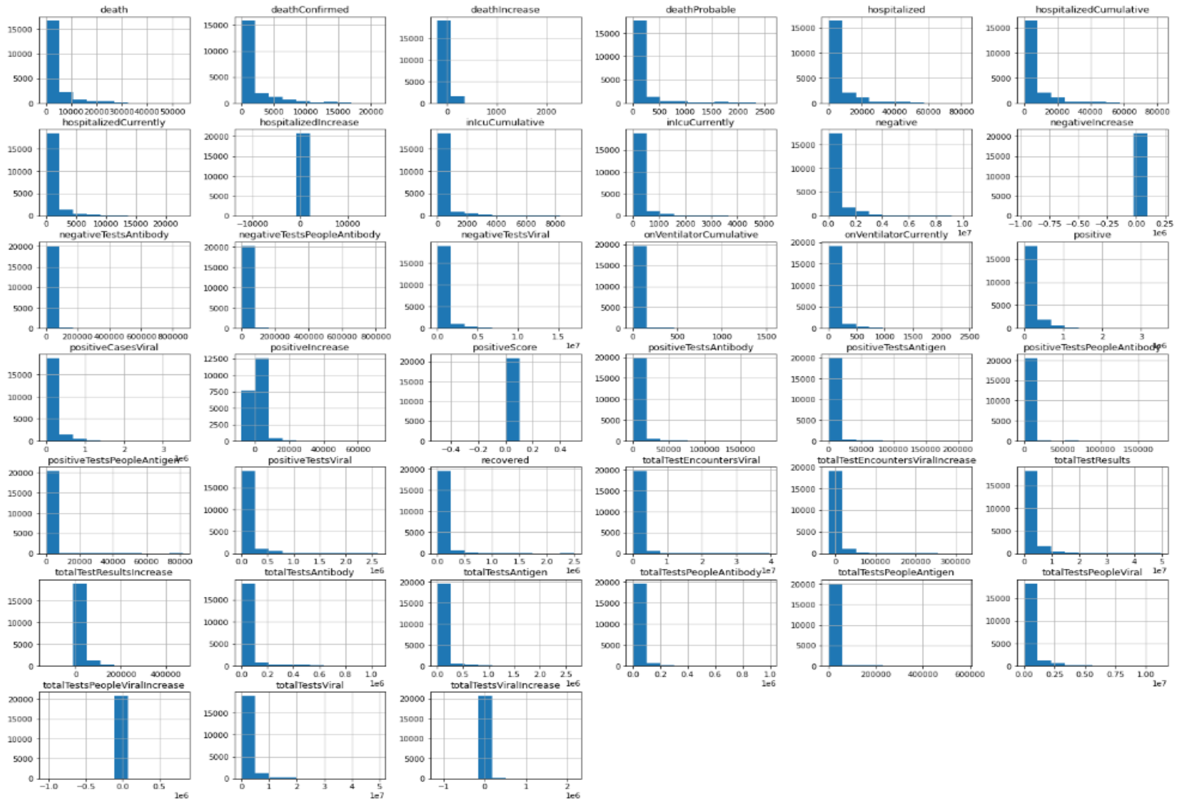


Fig. 12

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.

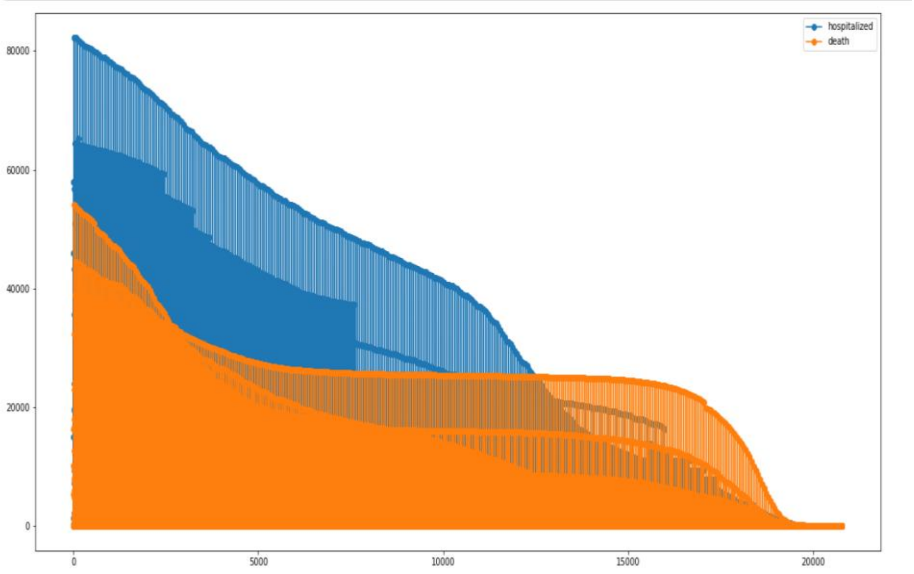
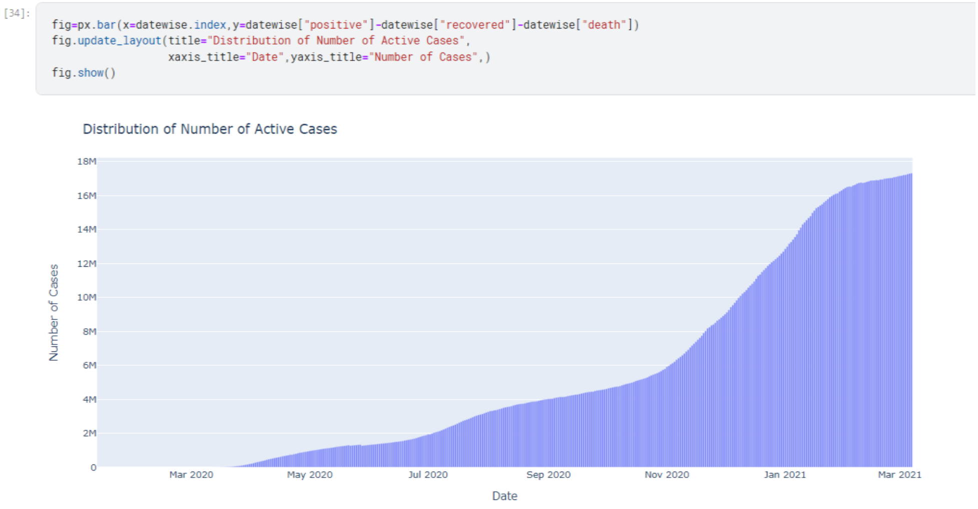


Fig. 13

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.

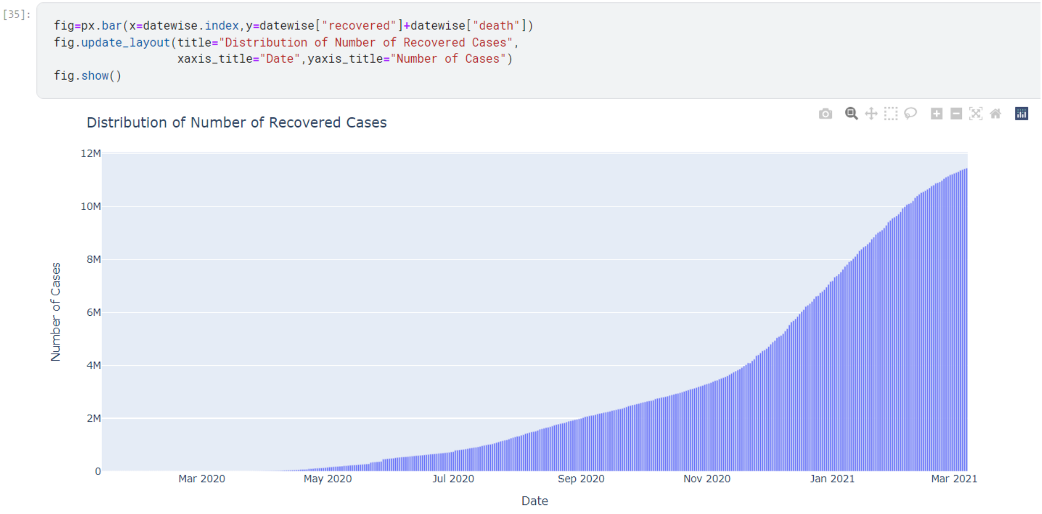
Even though the hospitalisations seems to be decreasing over time, the deaths are constant for a long period of time.



```
[34]:
fig=px.bar(x=datewise.index,y=datewise["positive"]-datewise["recovered"]-datewise["death"])
fig.update_layout(title="Distribution of Number of Active Cases",
                  xaxis_title="Date",yaxis_title="Number of Cases",)
fig.show()
```

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



```
[35]:
fig=px.bar(x=datewise.index,y=datewise["recovered"]+datewise["death"])
fig.update_layout(title="Distribution of Number of Recovered Cases",
                  xaxis_title="Date",yaxis_title="Number of Cases")
fig.show()
```

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.

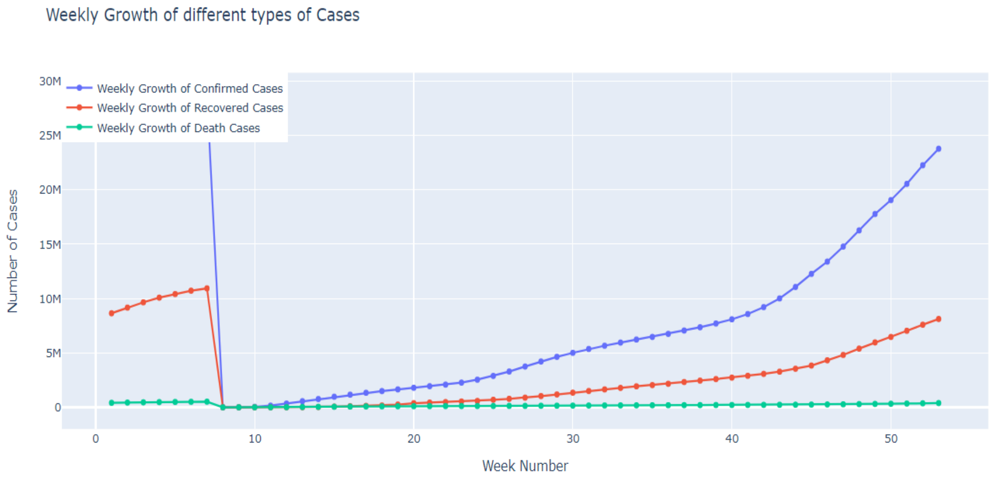


Fig. 16

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.

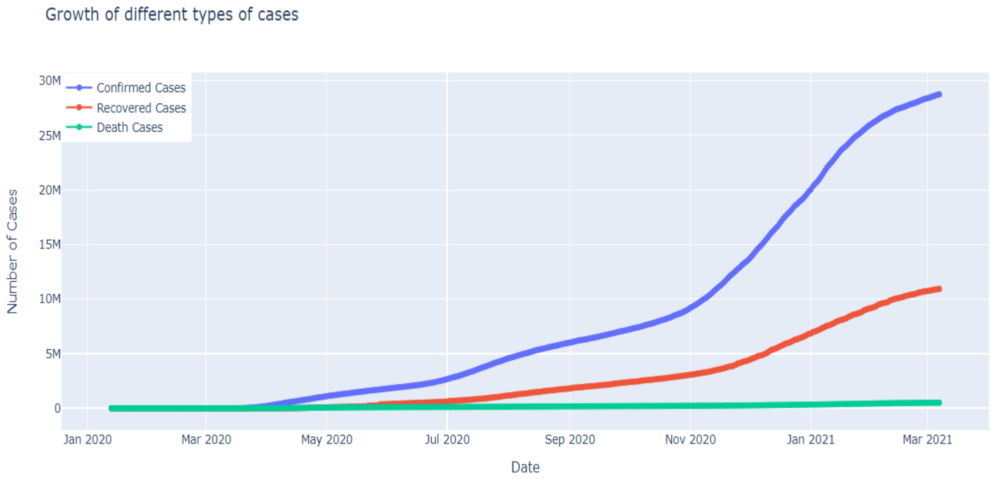
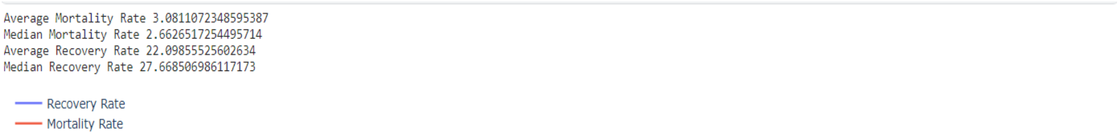


Fig. 17

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.



```
Average Mortality Rate 3.0811072348595387
Median Mortality Rate 2.6626517254495714
Average Recovery Rate 22.09855525602634
Median Recovery Rate 27.668506986117173
```

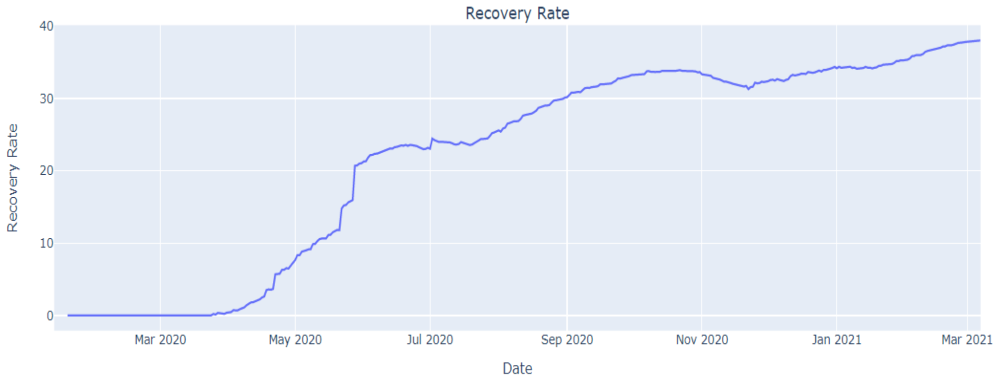


Fig. 18

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.

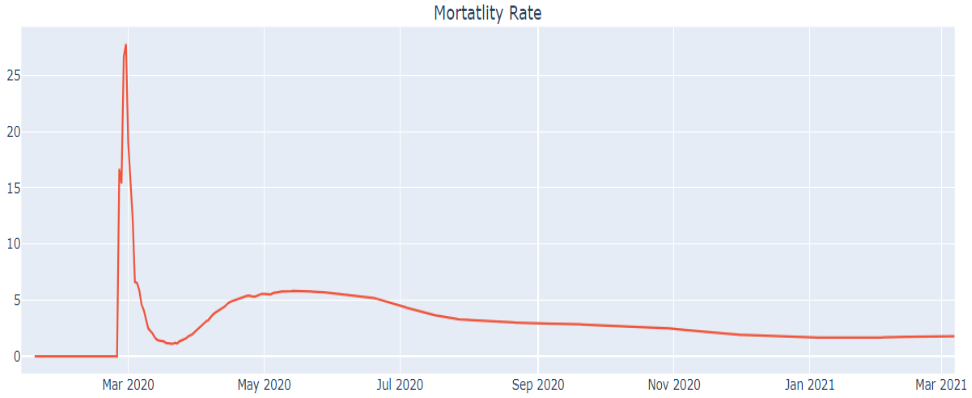
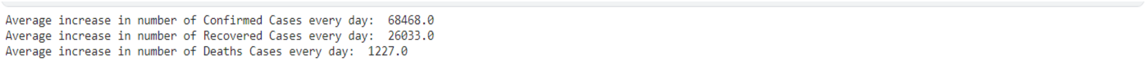


Fig. 19

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.



```
Average increase in number of Confirmed Cases every day:  68468.0
Average increase in number of Recovered Cases every day:  26033.0
Average increase in number of Deaths Cases every day:  1227.0
```

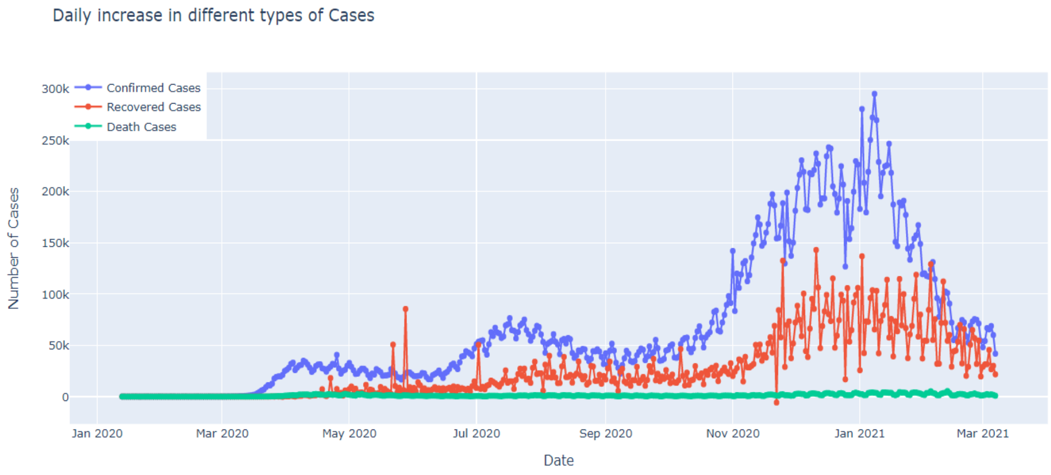


Fig.20

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.

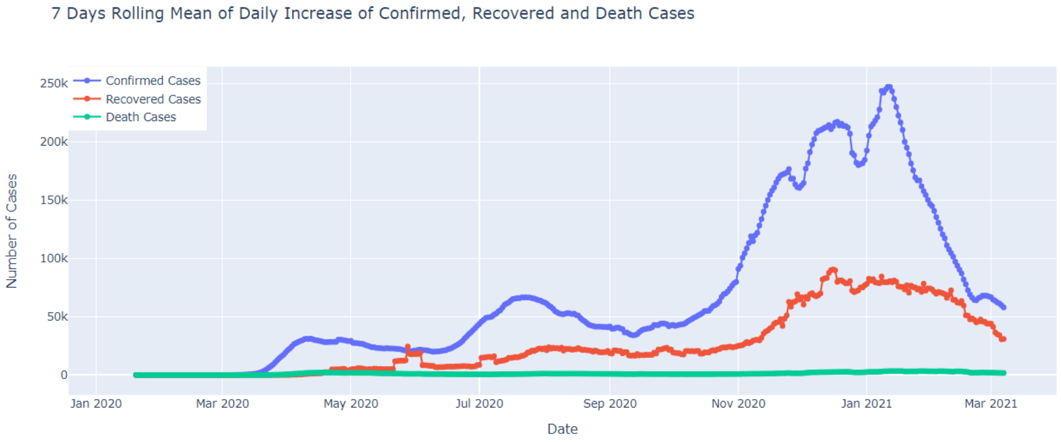


Fig. 21

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.

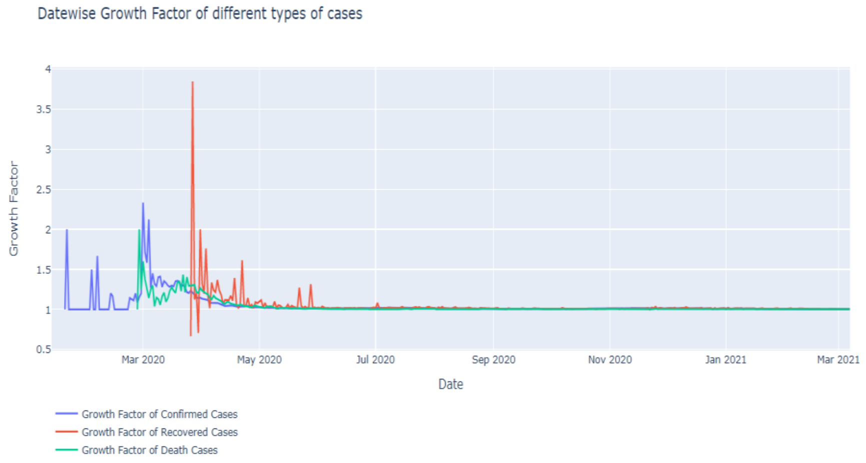


Fig. 22

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.

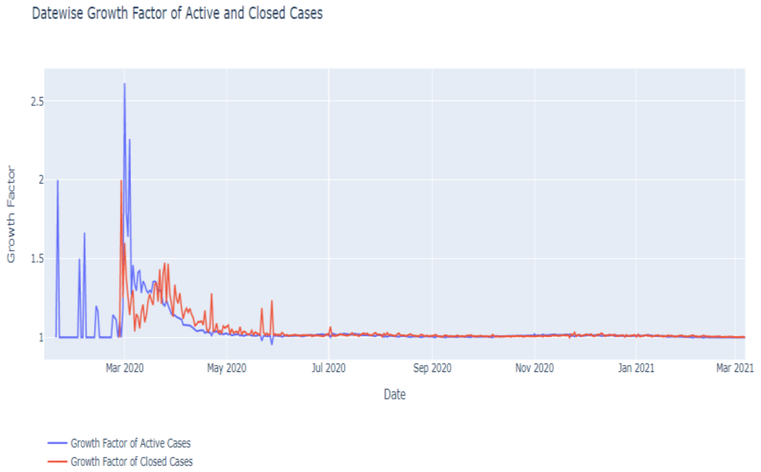
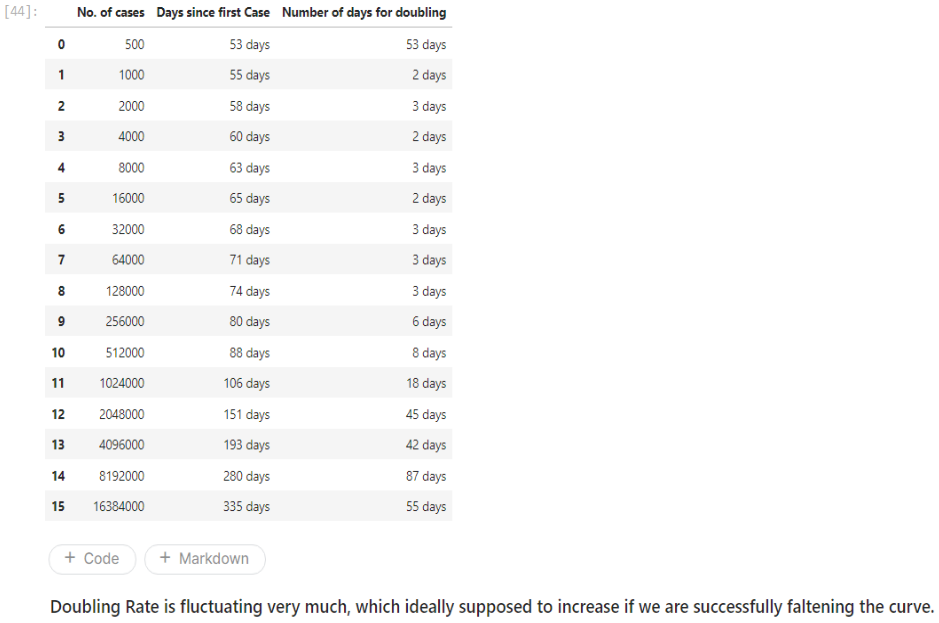


Fig. 23

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.



	No. of cases	Days since first Case	Number of days for doubling
0	500	53 days	53 days
1	1000	55 days	2 days
2	2000	58 days	3 days
3	4000	60 days	2 days
4	8000	63 days	3 days
5	16000	65 days	2 days
6	32000	68 days	3 days
7	64000	71 days	3 days
8	128000	74 days	3 days
9	256000	80 days	6 days
10	512000	88 days	8 days
11	1024000	106 days	18 days
12	2048000	151 days	45 days
13	4096000	193 days	42 days
14	8192000	280 days	87 days
15	16384000	335 days	55 days

Doubling Rate is fluctuating very much, which ideally supposed to increase if we are successfully faltening the curve.

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.

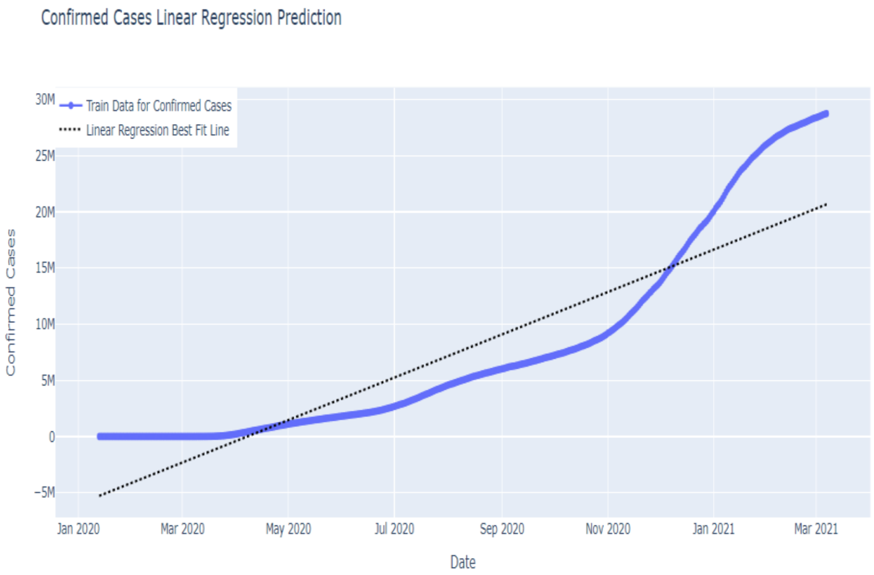


Fig. 25

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.

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

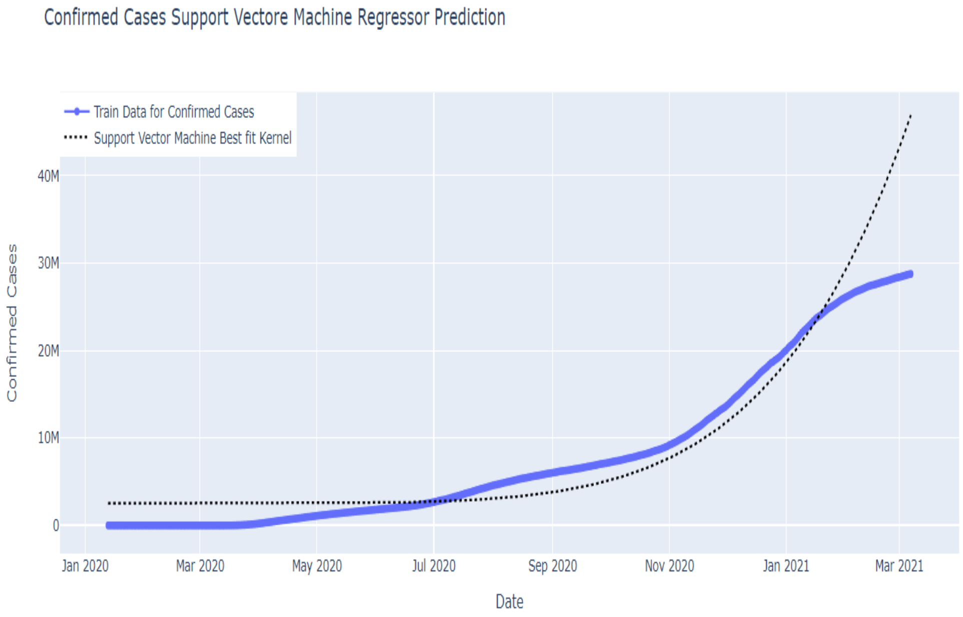


Fig. 26

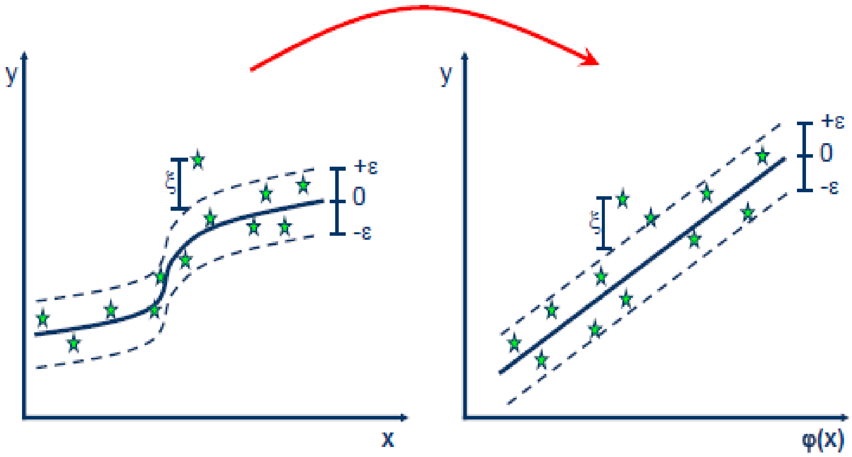
Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. First of all, because output 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

$$y = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) \cdot \langle \varphi(x_i), \varphi(x) \rangle + b$$

$$y = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) \cdot K(x_i, x) + b$$



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

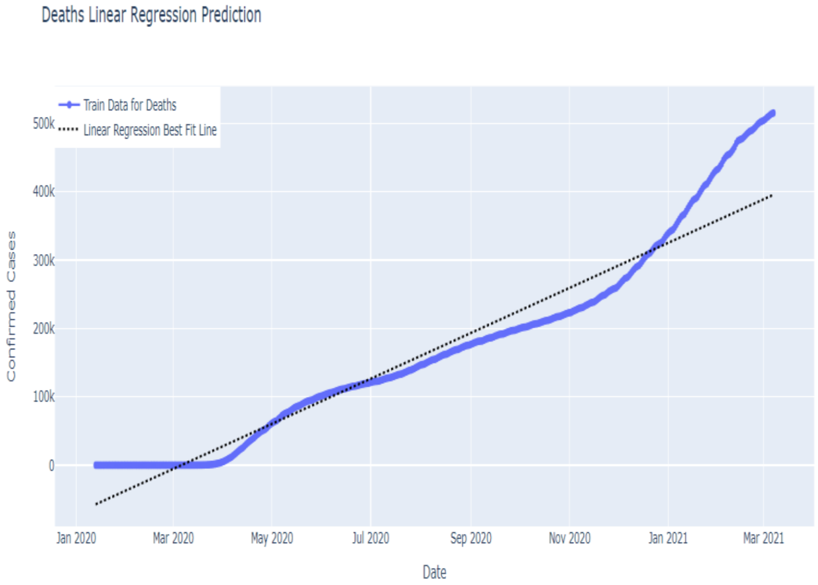


Fig. 27

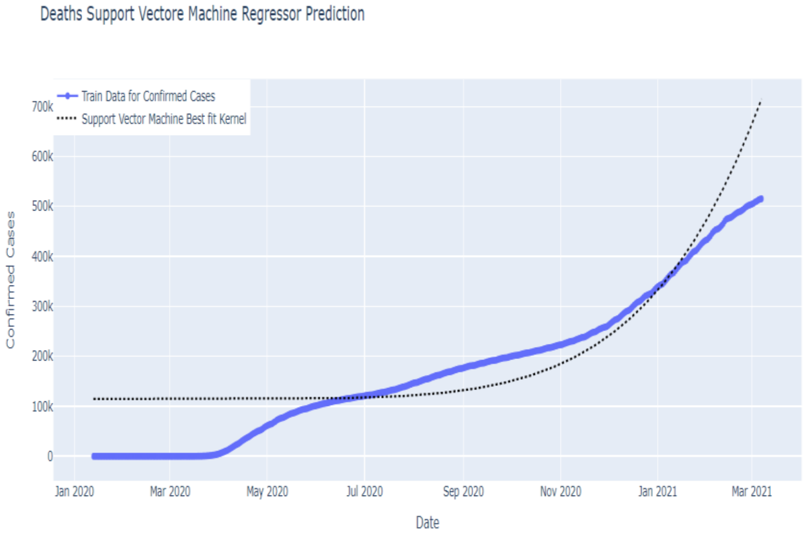

Deaths Support Vectore Machine Regressor Prediction
Train Data for Confirmed Cases
Support Vector Machine Best fit Kernel
Confirmed Cases
700k
600k
500k
400k
300k
200k
100k
0
Jan 2020
Mar 2020
May 2020
Jul 2020
Sep 2020
Nov 2020
Jan 2021
Mar 2021
Date

Fig. 28

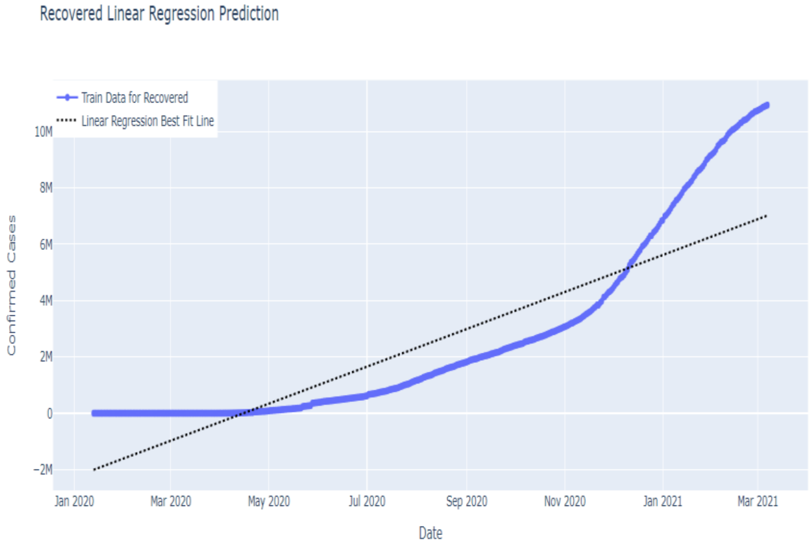

Recovered Linear Regression Prediction
Train Data for Recovered
Linear Regression Best Fit Line
Confirmed Cases
10M
8M
6M
4M
2M
0
−2M
Jan 2020
Mar 2020
May 2020
Jul 2020
Sep 2020
Nov 2020
Jan 2021
Mar 2021
Date

Fig. 29

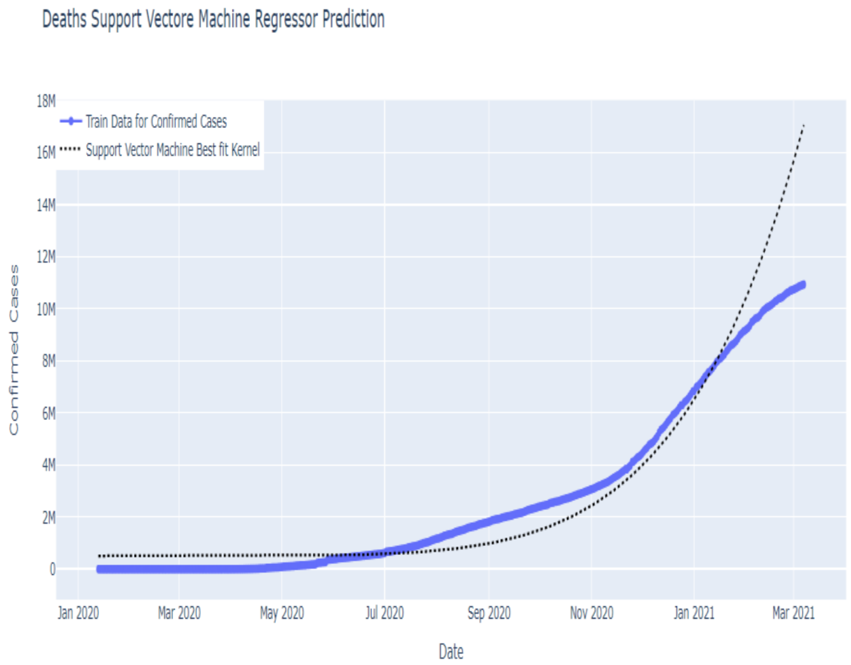


Fig. 30

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