Exploratory Analysis of Covid-19 in the United States for the period 2020 to 2022

TANISHQ CHERUKURI

Introduction

As of June 2020, the United States had the highest number of reported COVID-19 cases in the world. According to the Centers for Disease Control and Prevention (CDC), there had been over 2.2 million confirmed cases of COVID-19 in the country and more than 120,000 deaths. The US had seen a rapid increase in cases since the start of the pandemic and a corresponding increase in record numbers of new cases. The US also experienced a surge in hospitalizations due to COVID-19, with over 40,000 people hospitalized. As of June 2020, all 50 states had issued stay-at-home orders and social distancing guidelines in an effort to slow the spread of the virus. The states with the most confirmed cases are New York (over 150,000 cases), New Jersey (over 50,000 cases), California (over 34,000 cases), and Michigan (over 23,000 cases). The states with the most deaths are New York (over 11,000 deaths), New Jersey (over 2,400 deaths), Michigan (over 1,400 deaths), and Massachusetts (over 1,200 deaths).

The aim of this paper was to perform an EDA on Covid-19 data collected in the United States from the period July 2020 to July 2022.The goal was to get an in-depth understanding of the effects of Covid-19 on the lives of the people living the United States.

Data collection and description

The data was collected on Kaggle and transferred to Jupyter Notebook for analysis. The data contained a total of 15 columns and 60,060 rows. Out of the said 15 columns ,5 of them were of type object and the rest were either of type float or integer. Before the analysis all the rows with missing columns were dropped and the variable containing date information was confirmed into a date type variable for easy analysis.

Summary of the vital statistics

Total Cases	Deaths	Total Confirmed Cases	Total Confirmed death	Mortality rate %
19580497292	291026110	16678183421.0	257372973.0	1.540

Looking at the table above the total number of cases in the country from July 2020 to July 2022 was 19580497292, in said period the total number of reported deaths were 291026110. Both statistics show the extent to which Covid-19 impacted the lives of the majority of the people living in the United States. Additionally, the mortality rate which is a ratio between the total reported deaths in a certain population was 1.54%. A mortality rate of 1.54 is considered high. This means that out of every 1,000 people in the population, an average of 15.4 people died each year as a result of the virus. High mortality rates are typically associated with developing countries or regions experiencing famine or war.

Descriptive statistics

	Count	Mean	std	min	25%	50%	75%	Max
tot_cases	27977.0	699878.374	850345.736829	0.0	59514.0	360321.0	1006327.0	4885289.0
conf_cases	27977.0	596139.093577	750652.149053	0.0	52106.0	304171.0	869590.0	4640489.0
prob_cases	27977.0	103739.338099	152734.577584	0.0	31.0	27901.0	148134.0	763762.0

new_case	27977.0	1749.246881	4051.631016	-4803.0	29.0	535.0	1832.0	125572.0
pnew_case	27977.0	294.912392	783.572885	-6259.0	0.0	33.0	262.0	18072.0
tot_death	27977.0	10402.334418	12098.881484	0.0	1137.0	5154.0	16556.0	71408.0
conf_death	27977.0	9199.448583	11040.241549	0.0	1079.0	4686.0	14287.0	71408.0
prob_death	27977.0	1202.885835	1641.852397	0.0	0.0	319.0	1983.0	7889.0
new_death	27977.0	19.457626	46.617508	-352.0	0.0	5.0	20.0	1178.0
pnew_death	27977.0	2.372020	24.398251	-2594.0	0.0	0.0	2.0	1021.0

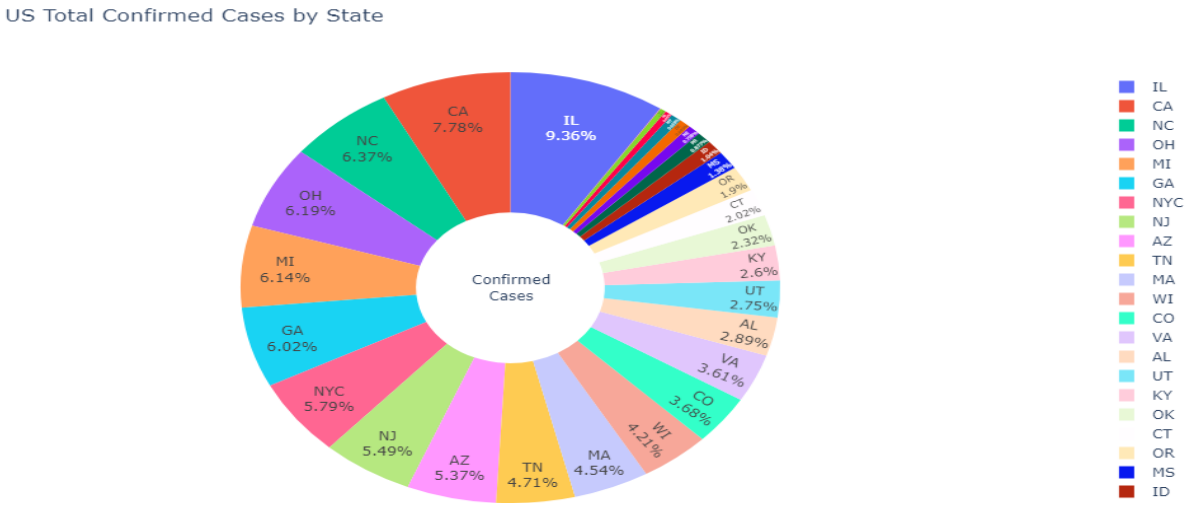
The table above is a descriptive statistics summary of the major vital statistics in regards to the effects of Covid-19 on the US population. As per the statistics the average number of confirmed Covid related cases were 699878.374 and ana average confirmed deaths of 10402. 334418.Both numbers are quite high and are a clearly indication of the effects of the virus in the country Additionally, other important statistics that were calculated include; the probability of new deaths which is a measure of the likelihood of death and the probability of new cases which is the likelihood of new cases. Both the said entities had high mean values of more than 1000 which is quite significant.

Exploratory analysis

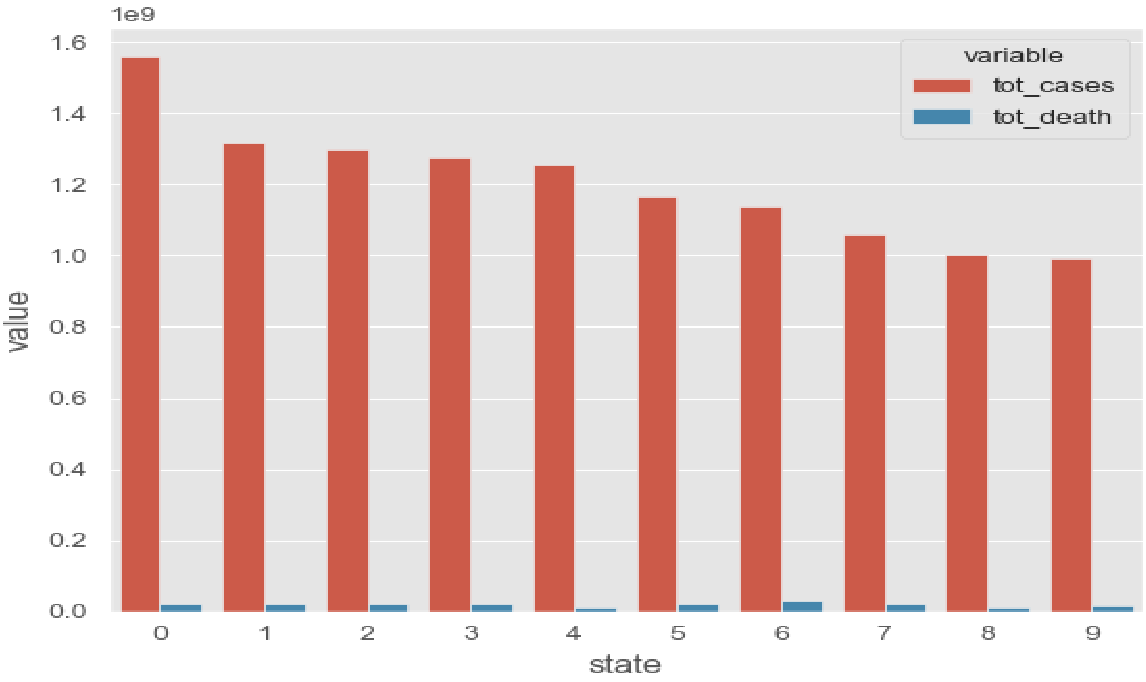
Exploratory data analysis (EDA) is a type of data analysis used to gain insights and identify patterns in datasets. It involves summarizing and visualizing the data in various ways to understand the underlying structure and relationships in the data. With the outbreak of the novel coronavirus (COVID-19) pandemic, EDA is essential for understanding the spread of the virus and the impact it has had on different countries and regions. By exploring the data, researchers can identify trends and patterns that can help inform public health measures and policy decisions. Additionally, EDA can be used to identify areas where more research and data collection is needed to better understand the virus and its impact on society.

Total confirmed cases by state

The aim of this EDA was to get a deeper understanding of the effects of the virus per state in regard to the total number of confirmed cases. To do this a pie chart was created showing the exact percentage of confirmed cases per county and the result are as seen in the image below.



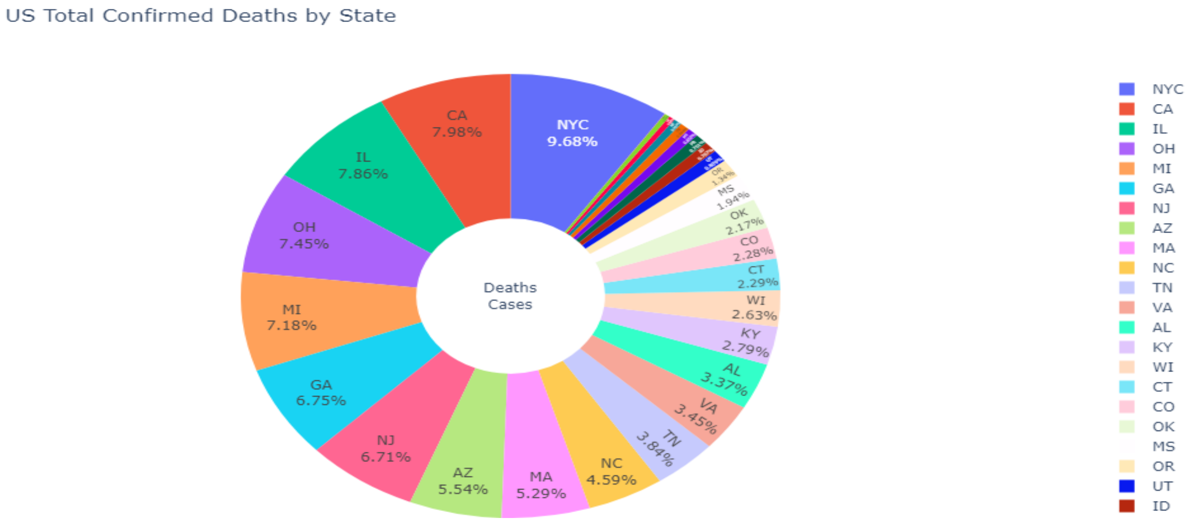
Looking at the chart above the state with the highest confirmed death cases was Illinois, with a confirmed death cases of 9.98% followed closely by California and North Carolina. The same information was represented in a barograph as seen in the image below.



The information on the bar graph is similar to that on the chart , where the highest number of cases recorded per state was in Illinois followed by California and New York City.

Total confirmed deaths by state

The aim of this EDA was to get an in-depth understanding of the States that were affected mostly by the virus in regard to the total number of confirmed deaths. The results of the EDA were presented in a pie chart as seen in the image below.



The results show that the state with the highest number of recorded confirmed deaths was Ney York City whose total number of confirmed deaths were 9.86% of the entire population followed closely by California and Illinois.

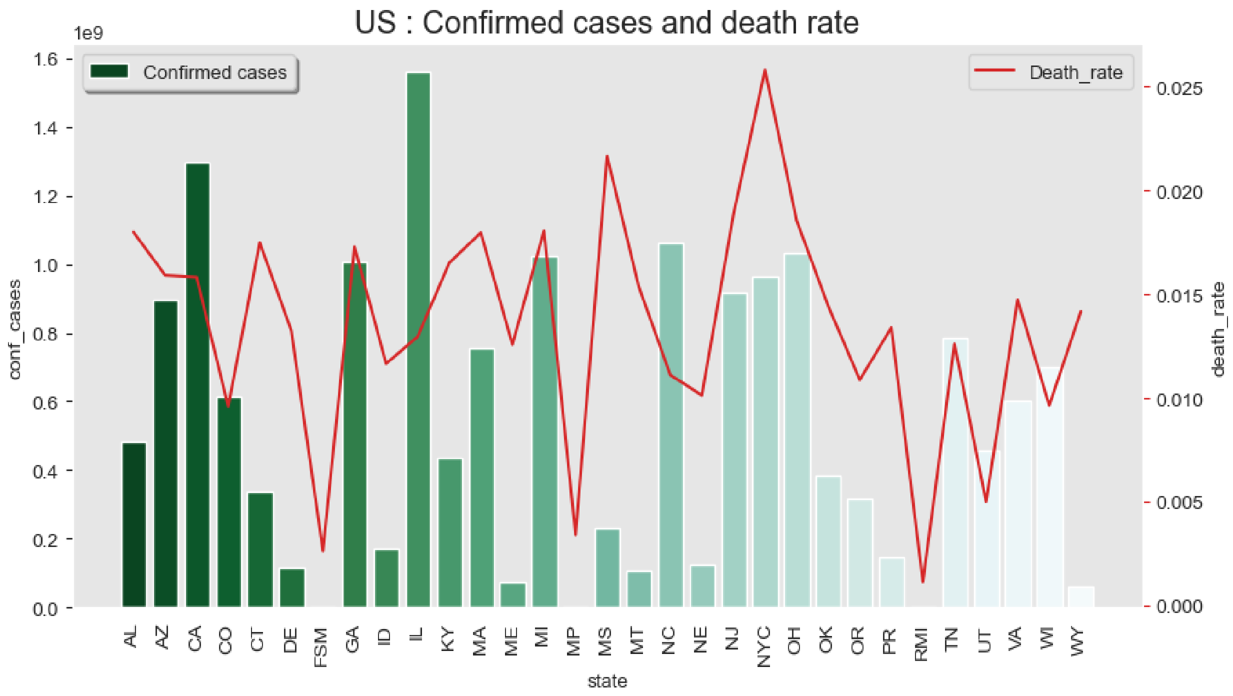
Death rate by state

Death rate is the number of deaths per unit of time, usually expressed per 1,000 or 100,000 individuals in a population. It is usually measured as the number of deaths per 1,000 individuals per year. It can also be expressed as the number of deaths in a given period of time, divided by the size of the population at the beginning of that period. The death rate by state in regard to Covid-19 in the US are as seen in the table below.

State	Confirmed_cases	Confirmed_deaths	Death_rate
NYC	9.648760e+08	24910276.0	0.025817
MS	2.309648e+08	5002470.0	0.021659
NJ	9.162550e+08	17265102.0	0.018843
OH	1.032921e+09	19165286.0	0.018554
MI	1.023664e+09	18483885.0	0.018057

The table contains information for the top 5 states with the highest death rate. New York City had the highest death rate followed by Mississippi. In both the states the death rate was a value of about 0.02, which implies 20 out of 1000 people died due to the virus.

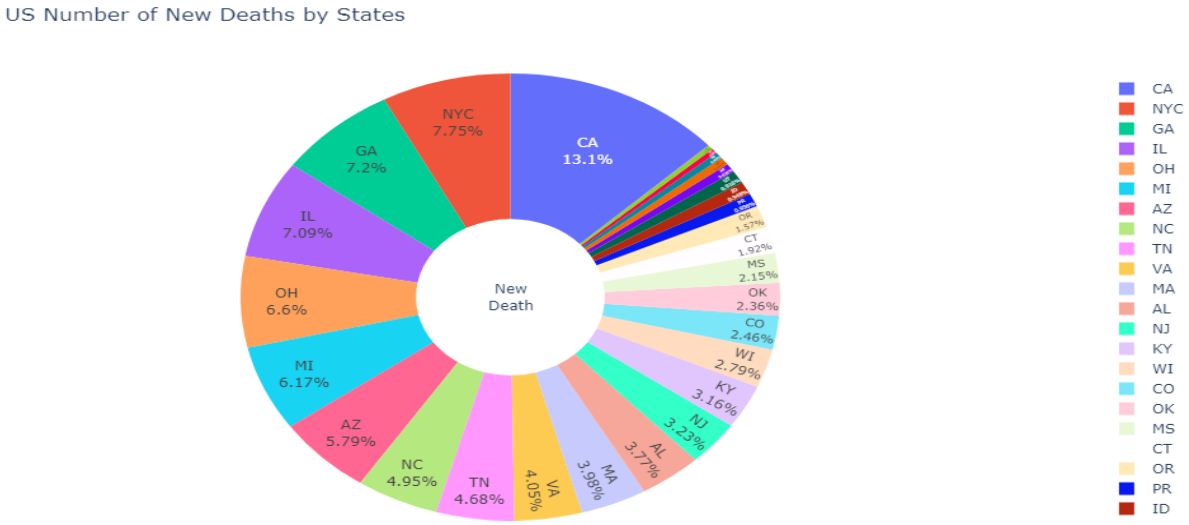
Confirmed cases and death rates per state



As per the graph Illinois had the highest number of confirmed cases, and a death rate of above 0.015 While New York City had the highest death rate of 0.025 and over 1 million total number of confirmed cases.

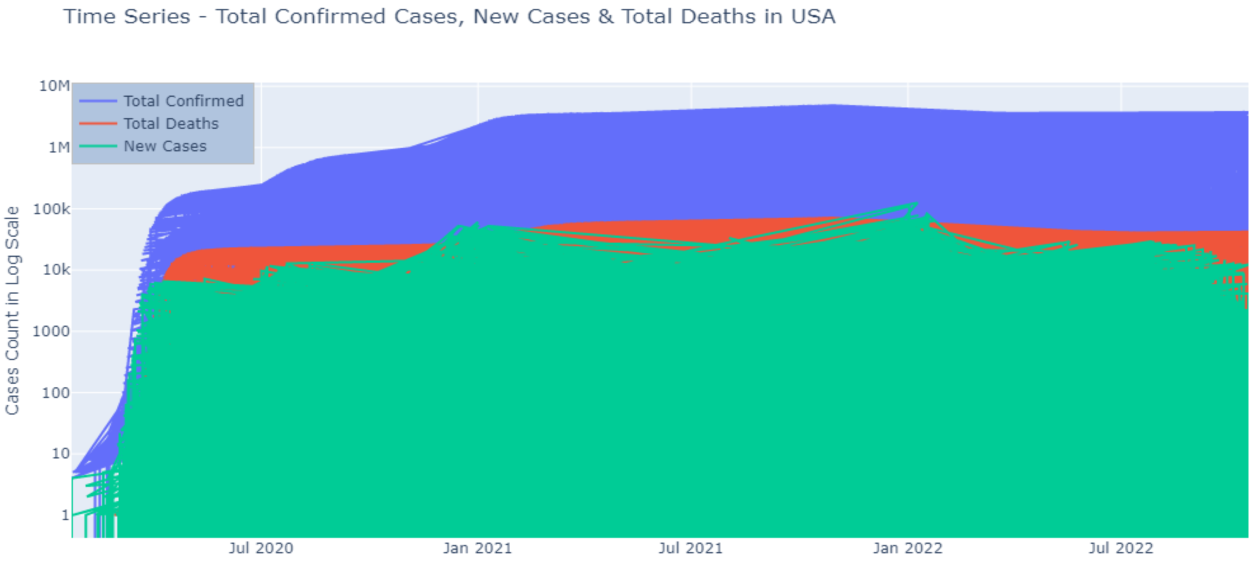
Number of new deaths per state

This EDA aims at understanding the total number of new deaths recorded per state. The goals is to be able to know which state was mostly afflicted with new covid related deaths. The results of the EDA were presented in a pie chart as seen in the image below. As seen in the image , California had the highest number of new reported death cases .As the number of new deaths were 13.1% of the entire population followed closely by Ney York City, Georgia and Illinois.



The trend of confirmed cases, deaths and new cases.

The aim of this EDA was to get a clear picture of the trend of vital statistics in regards to the virus between the period July 2020 and July 2022.By analyzing trend we get to have a clear picture of the effects of the virus in the 2-year period. The results are as seen the image below.



Looking at the results there was a continuous spike in the total number of confirmed cases, total deaths and new cases from early 2020 to January 2021.But from January 2021 to July 2020 there was no significant increase in the said variables. And so, the assumption is that by that time the virus had reached its saturation point and also the government had put in enough measures to curb it. Additionally, as from January 2022 to July 2022 there was a decrease in the number of newly recorded cases which also signifies better government measures to decrease its effects.

Conclusion

The aim of this paper was to perform an EDA on various vital statistics in regards to the effects of Covid-19 on the US population for the period 2020 to 2021. The results of the analysis showed that the country was greatly impacted by the virus and state wise New York City, Illinois, California and Georgia were some of the states that were greatly impacted by the virus. The reason could be tied to the fact that the said states are densely populated. Additionally, it could be due to the fact that each state had its own Covid-19 policies that were somewhat in tandem with ones implemented by the national government with a slight variation. But the effects of the virus were only significant in the first year .2020 was a bad year for the country in regard to the virus , but the implemented government polices and the WHO regulations reduced the rise of the virus and by early 2021 there was a decrease in the total number of newly recorded cases.