Name: Refilwe Tlhopho Matema

Date:12/04/2025

**DATA SCIENCE & DATA ANALYTICS|AI|ML**

**PROJECT 3 REPORT**

**Introduction**

The main objective of this project is to understand how the cases of covid-19 are progressing within each region individually and then on a global scale as a whole.The progression of COVID-19 cases and deaths at the county and national levels in the US will be analyzed. The technician will identify regions with rapid case increases, high slopes of cases per time, and windows of peak fatalities.The aim is to gain insights into the pandemic's spread and inform strategies like lockdown planning. Two datasets will be used for this project namely; time\_series\_covid19\_confirmed\_US.csv and time\_series\_covid19\_deaths\_US.csv.

**Columns:**

* UID: Unique identifier
* Iso2:Country codes
* Iso3: Country codes
* code3: Country codes- abbreviations for countries
* FIPS: Federal Information Processing Standard code. Federal Information Processing System (FIPS) Codes for States and Counties.FIPS codes are numbers which uniquely identify geographic areas. The number of digits in FIPS codes vary depending on the level of geography.
* Admin2: County-level administrative unit
* Province\_State: State or province
* Country\_Region: Country
* Latitude, Longitude: Geographic coordinates
* Combined\_Key: Unique identifier combining location and admin 2
* Time series data: Daily counts of confirmed cases or deaths

**10 Research Questions were answered with visualizations:**

1. **What were the trends in COVID-19 confirmed cases in the US?**

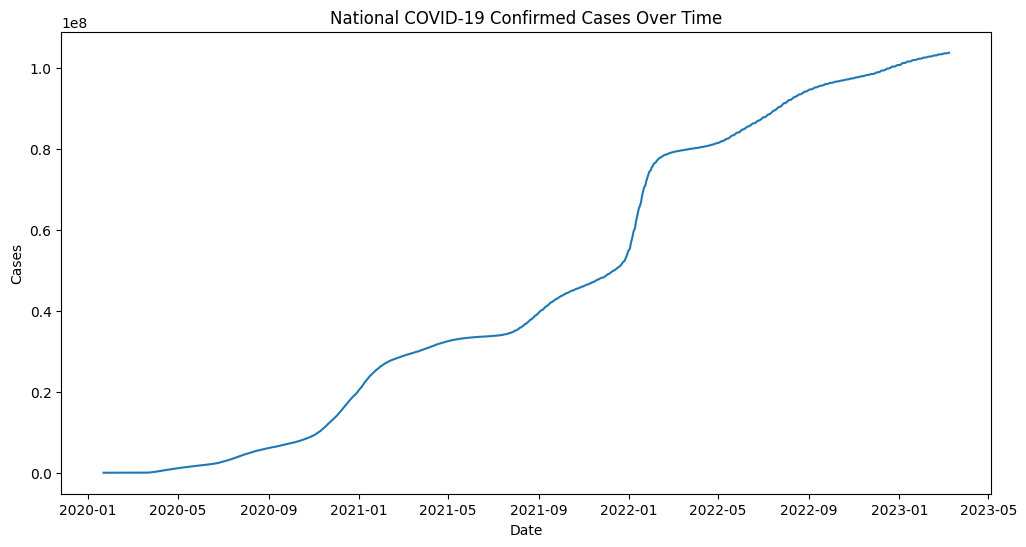


Figure 1: National Covid 19 confirmed cases over time

Overall, the number of Covid 19 confirmed cases increased from 2020 to 2023. Jan 2022 to March 2022 is where we see the largest increase in the number of confirmed cases. We see another steep increase in Jan 2021 to March 2021. The number of cases were still generally low in the first 6 months of 2020. The significant rises in cases in these particular times could be due to a new Covid 19 variant which could also be an appropriate time to initiate lockdown.

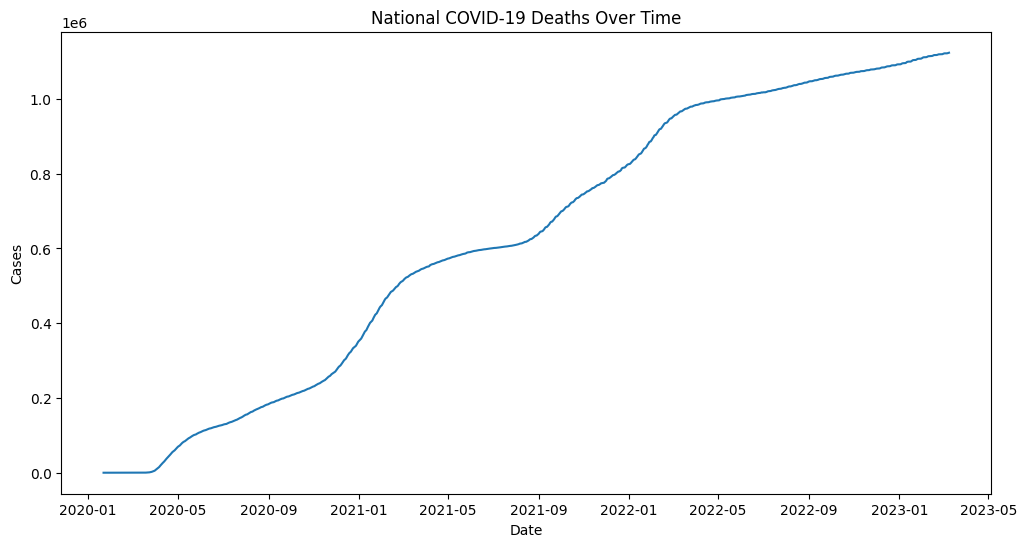
1. **What were the trends in COVID-19 deaths in the US?**

Figure 2: National Covid 19 deaths over time

There is a general increase in the amount of Covid 19 deaths over the years. There was a significant increase in May 2020, Jan 2021 to March 2021 January to May 2025.These significant increases show peak fatality times.

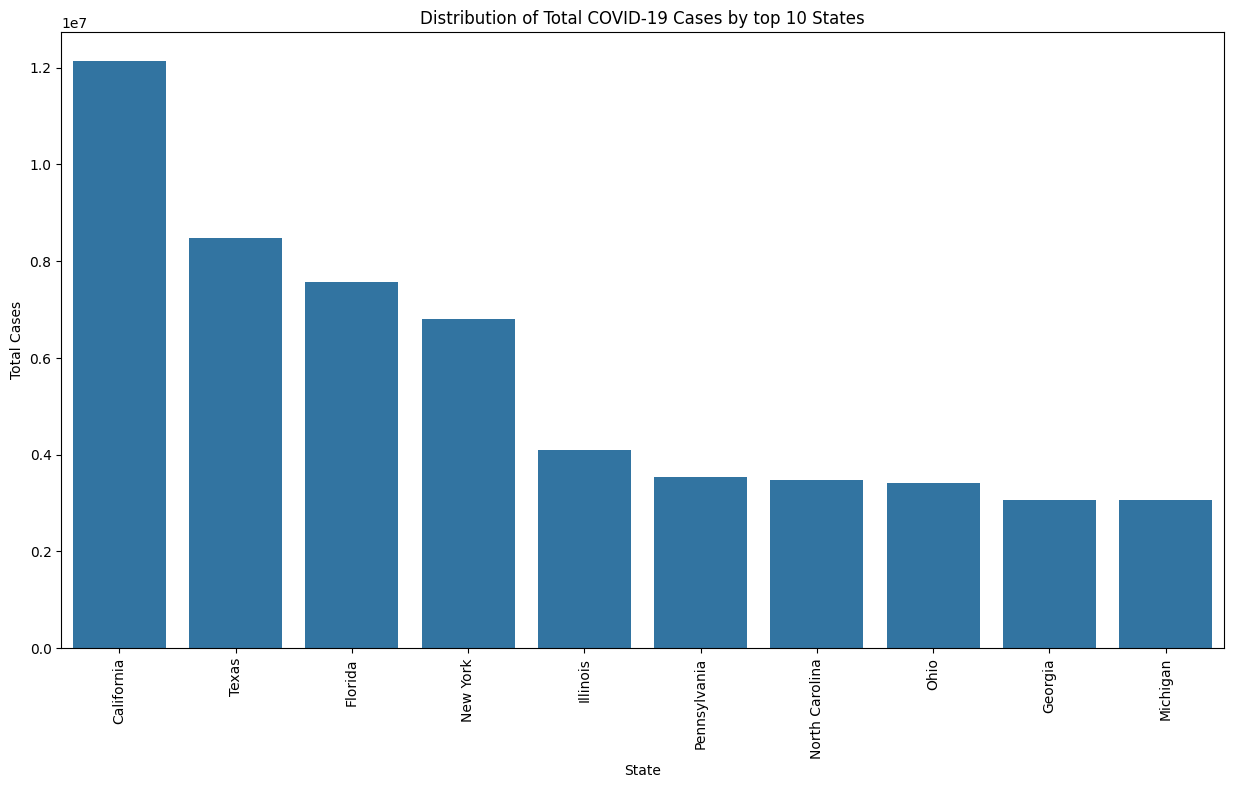
1. **Which US state exhibited the highest number of increases in COVID-19 confirmed cases?**

Figure 3:Distribution of total Covid 19 cases by 10 states

Carlifonia is the city with the highest number followed by Texas than Florida. This is where most people had Covid 19 cases. This could be due to a large population in California or the type of popular activities that are popular. This will help to choose which state city to implement lockdown first.

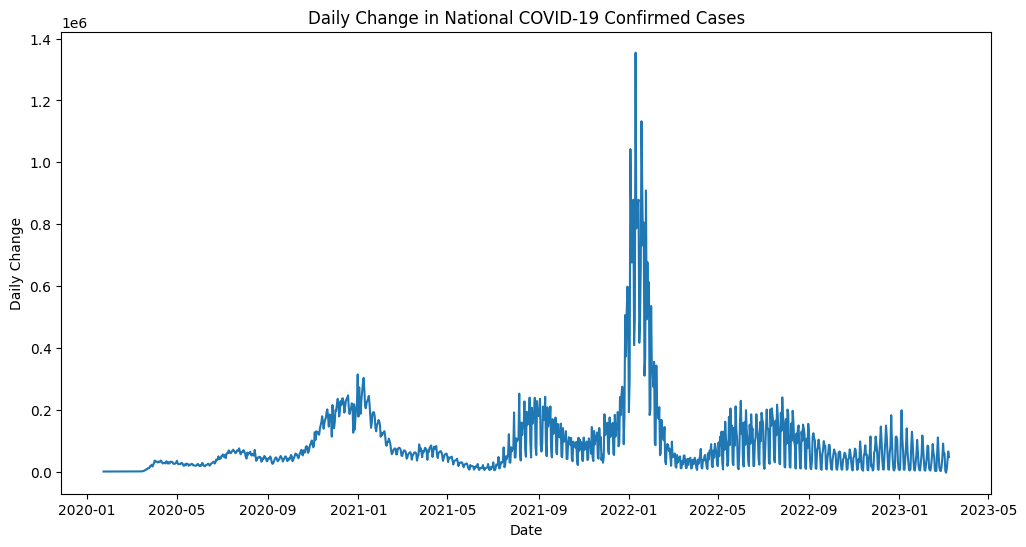
1. **When were the windows of peak COVID-19 confirmed cases in the USA?**

Figure: 4:Daily change in National Covid 19 confirmed cases

Daily change refers to how much the total number went up or down from one day to the next. For the first 9 months of 2020 the daily change in confirmed cases was low. There were different peaks in confirmed cases in January 2021, September 2021, January 2022 which was the highest peak of cases , June 2022 and January 2023. These peaks could represent when there new Covid 19 variants which would also be the best time to have a lockdown.

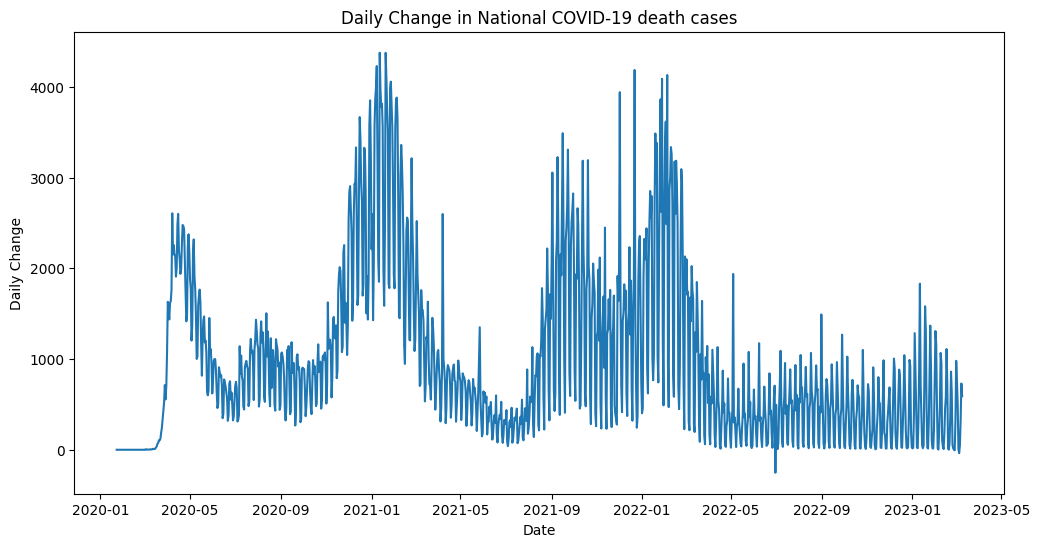
1. **When were the windows of peak COVID-19 fatalities in the USA?**

Figure 5: Daily change in National Covid 19 death cases

Daily change refers to how much the total number went up or down from one day to the next. It's the difference between each day's total and the previous day's total.The highest number of daily increase in Covid 19 deaths is seen in the period of January 2022 to March 2022. The first 10-11 months of 2020 there was a generally very low death rate. There are other times when the death daily changes peaked which are Jan 2021 Sep 2021 and June 2022 - July 2022 and Jan 2023.

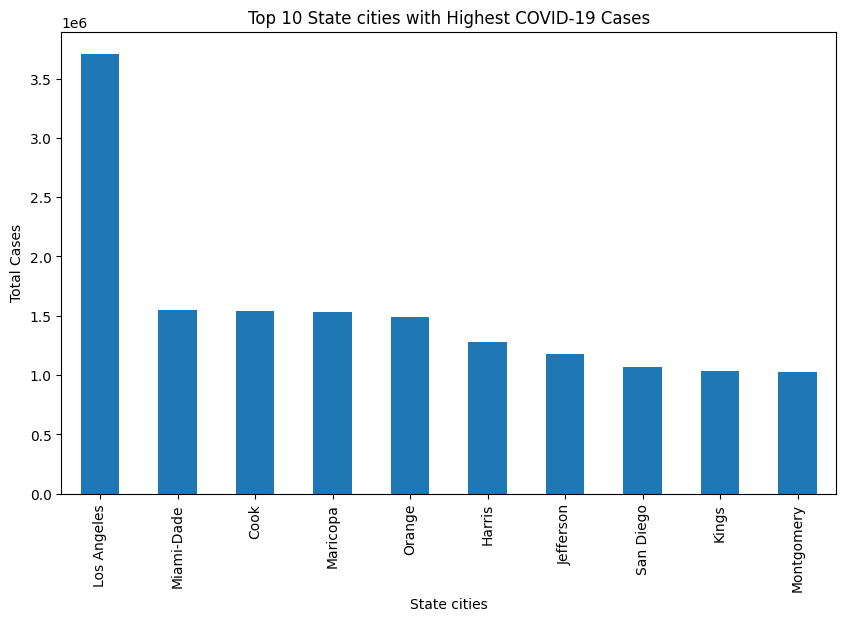
1. **Which state cities had the highest number of Covid 19 confirmed cases?**

Figure 6:Top 10 state cities with highest Covid 19 cases

Los Angeles is the city with the highest number followed by Manti-Dade than Cook. These are states where most people had Covid 19 cases. This could be due to a large population in California or the type of popular activities that are popular. This will help to choose which state city to implement lockdown first.

1. **What is the relationship between the progression of COVID-19 cases and deaths?**

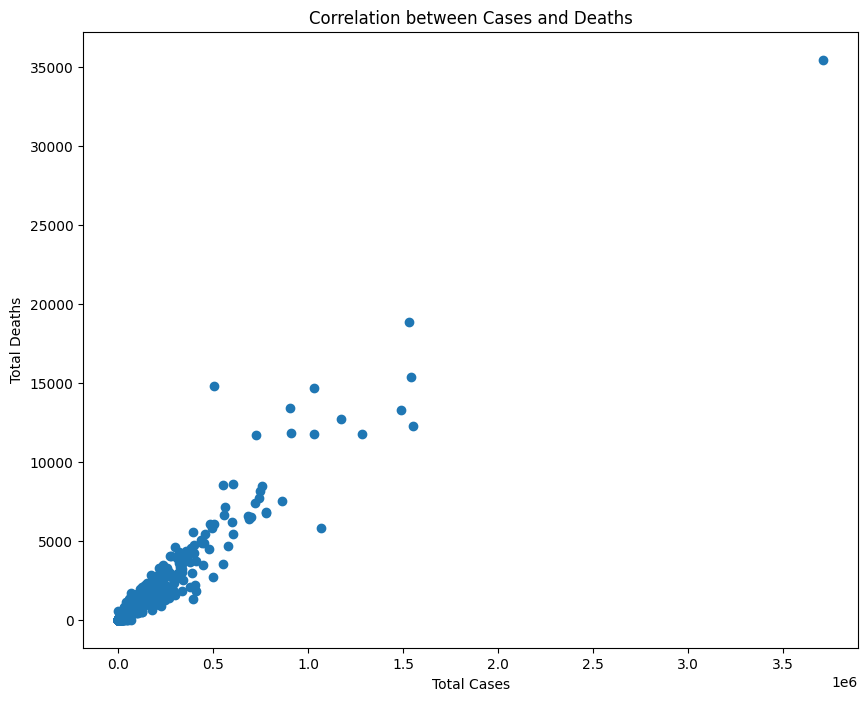


Figure 7:Relationship between cases and deaths

There is a positive correlation between total cases and total deaths this means that for the number of total deaths increases as the total number of cases increases. Most of the total number of deaths are under 10000. This means it is possible to predict when to do a lockdown.

1. **What is the total overall count of Covid 19 confirmed cases and total deaths? Are there more cases than death?**

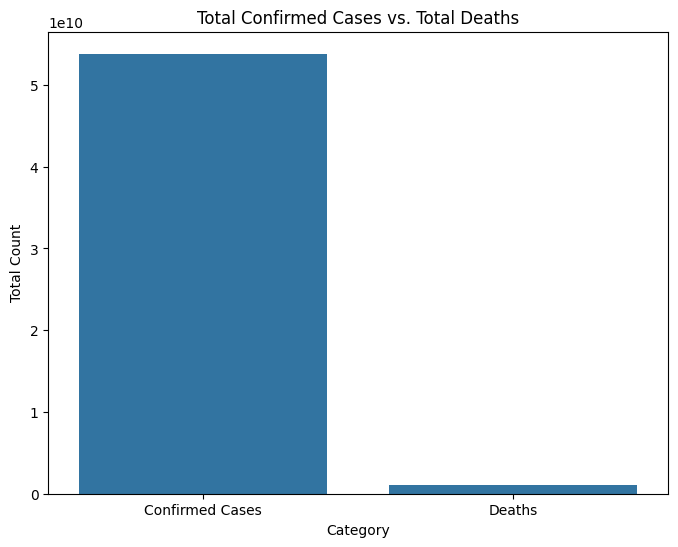


Figure 8: Total confirmed cases vs total deaths

Overall there was a higher count of confirmed cases than deaths in the Us. This means that most people who had Covid 19 disease did not die. Due to the low death rate it may have not been necessary to have a lockdown but still continue with strict measures like social distancing to slow down the spread of the disease.

**9. How do the confirmed cases and deaths compare in the US States?**

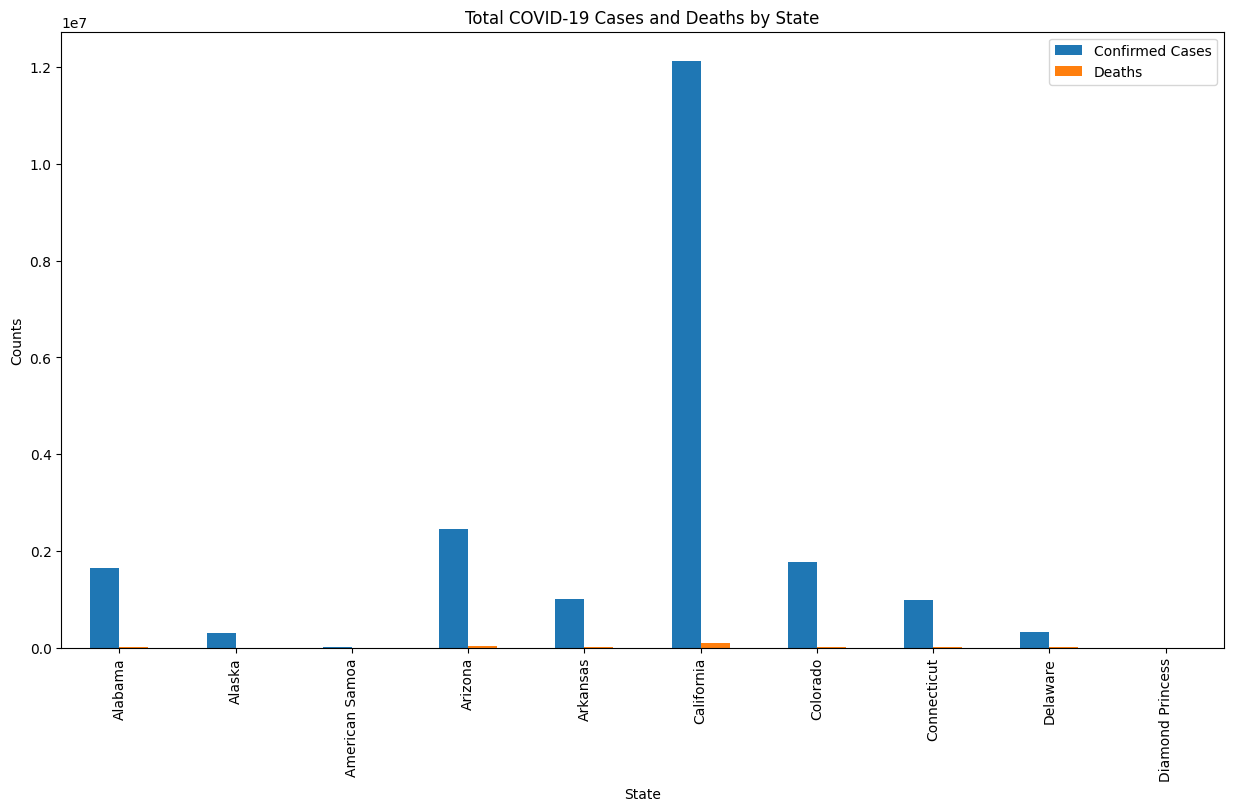


Figure 9: Total confirmed cases and deaths in US states

In terms of Covid19 confirmed cases California which is a US state had the highest count. The cases are significantly higher than all the other states. Overall the death count was very low for all the states however the California state had the highest death count. The confirmed cases for Carlifonia are significantly higher than the number of deaths.

**`10. What are the trends for regions according to longitudes and latitudes with Covid 19 deaths?**

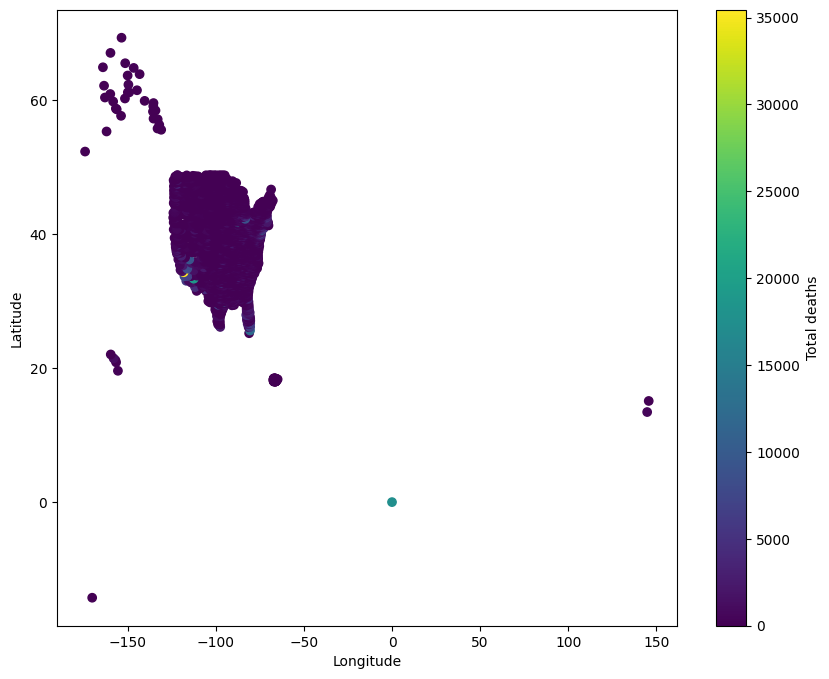
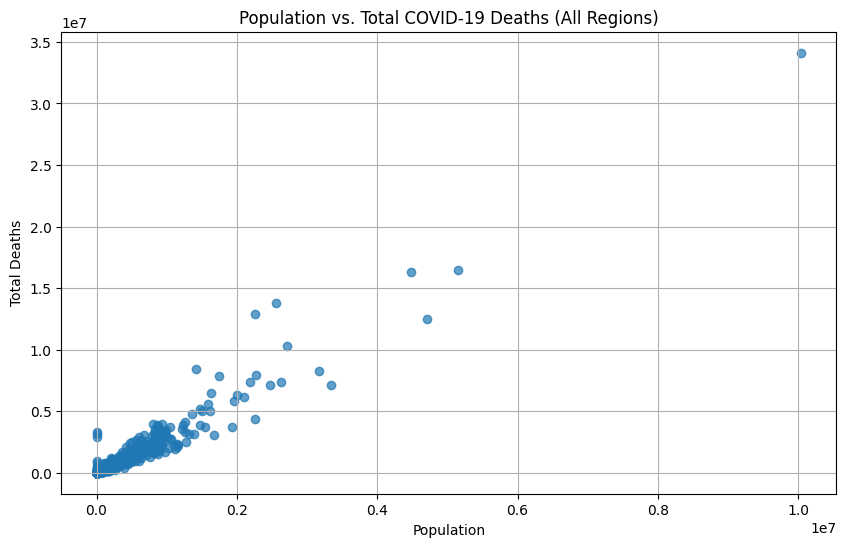


Figure 10: Relationship between location with Covid 19 deaths

There are certain regions within very little death rates from a longitude -50 to -150 and a latitude of 25 to 50. Which means that the geographical area or location of these state cities or states could be a factor in the spread of the Covid 19 disease and deaths.



**Figure 11: Relationship between total deaths and populatiom**

As the population increases the total deaths increases up to 4 million population and total deaths up to 15 million death cases.

**Findings from descriptive statistics for the covid 19 confirmed cases** dataset indicate that the mean latitude (36.72) and longitude (-88.64) place the data's center roughly in the central-southern United States. The standard deviation of latitude (9.08) and longitude (21.78) show a wide spread of data points across the US. The mean, 25%, and 50% are mostly 0 in the early days. This indicates that the virus was just starting, and most locations had no reported cases. The std (standard deviation) also increases, which means the spread of cases across different locations becomes more varied. Some areas were hit much harder than others. There's a clear upward trend in confirmed cases from the beginning of 2020 to early 2023.The descriptive statistics show a pandemic that started slowly, then rapidly increased, and had a highly variable impact across different regions.

**Findings from descriptive statistics for the covid 19 confirmed deaths** dataset show that death rates lagged behind confirmed case rates, as expected.Similar to the confirmed cases, the mean, 25%, and 50% are mostly 0. This indicates minimal deaths reported in most locations during the early stages of the pandemic.The std also increases, showing a growing disparity in death counts across regions.The death data's latitude and longitude distribution mirrors the confirmed cases data, suggesting that the same areas were heavily impacted by both cases and deaths.

The 25 and 50 percentile values are very low, compared to the 75 percentile and max values. This shows that the majority of the locations had a much lower death count, than the few locations with high death counts.The death statistics reveal a pattern of concentrated fatalities in specific regions, highlighting the need for targeted public health interventions.

**The machine learning model**

Recurrent Neural Network (RNN) models are particularly well-suited for time series forecasting, such as predicting COVID-19 confirmed cases, due to their inherent ability to capture temporal dependencies within sequential data. Unlike traditional models that treat data points as independent, RNNs possess a "memory" that allows them to consider past observations when making predictions. The data, sourced from the c19 confirmed and c19 deaths dataset is suitable for RNN modeling since it is non-stationary. The steps taken included:

* **Time Series Aggregation:** Daily confirmed case counts were aggregated either at the national level or for a specific county, depending on the desired scope of the forecast.
* **Date Indexing:** The time series index was converted to datetime format to facilitate time-based analysis.
* **Data Scaling:** RNNs often perform better when the input data is scaled to a specific range (e.g., 0 to 1 or -1 to 1).MinMaxScaler from scikit-learn is used for this.
* **Stationarity Check:** An Augmented Dickey-Fuller (ADF) test was performed to assess the stationarity of the time series.Dickey-Fuller Test Results: The test results confirm that the time series data for Covid 19 confirmed cases and deaths are non-stationary.
* **Train-Test Split:** The time series data was divided into training (80%) and testing (20%) sets.
* **Model Training:** An RNN model was instantiated and trained on the preprocessed training data.

ARIMA performs well with linear trends and seasonality however RNNs excel at capturing complex, non-linear patterns and long-term dependencies. RNN was chosen because it handles nonstationary data much better.

Metrics evaluation

For general evaluation, Mean absolute error (MAE), root mean squared error(RMSE), and Mean absolute percentage error (MAPE) are commonly used.

The MAE is 0.02% which means the model's predictions on the test dataset deviated from the actual values by only 0.02%. The model's predictions are remarkably close to the actual values which means that it performs exceptionally well on the test data.

Inferential statistics

There is a statistically significant **increasing trend in the daily confirmed COVID-19 cases**.A positive Mann-Kendall statistic indicates an increasing trend.The very large number indicates a very strong increasing trend.This trend could be due to several factors, such as the emergence of new, more transmissible variants, changes in testing capacity or reporting practices, relaxation of public health measures and seasonal variations in virus transmission.

This indicates that there is a consistent pattern of **increase or decrease in the daily death counts over the period of time analyzed**. A positive Mann-Kendall statistic indicates an increasing trend in the daily COVID-19 death counts over time. This result strongly suggests that, overall, the daily death counts increased over the time period examined.This could be due to various factors, such as waves of new COVID-19 variants, changes in public health policy, seasonal effects and the impact of vaccination rates over time.

There is a **statistically significant difference in mean death rates between densely and sparsely populated regions**. The extremely low p-value provides very strong evidence that there is a real and substantial difference in COVID-19 death rates between densely and sparsely populated regions.There is a real and substantial difference in COVID-19 death rates between densely and sparsely populated regions.This shows that the **dense regions were very heavily impacted by the virus.** The results highlight the vulnerability of densely populated areas to infectious diseases like COVID-19.This has implications for public health policies, resource allocation, and targeted interventions.

Conclusion

The exploratory data analysis reveals distinct patterns in the COVID-19 pandemic's progression across the United States, highlighting critical periods for potential intervention. Notably, both confirmed cases and deaths exhibited significant surges in early 2021 and early 2022, suggesting the likely emergence of new, highly transmissible variants during these times. These periods, characterized by steep increases in daily case changes and coinciding with peaks in fatality rates, should be considered prime windows for initiating stringent lockdown measures.

The inferential statistics show that there is a statistically significant trend in the daily confirmed COVID-19 cases and that there is a statistically significant difference in mean death rates between densely and sparsely populated regions.The dense regions were very heavily impacted by the virus. Furthermore, there is an increase or decrease in the daily death counts over the period of time analyzed.

Geographically, the central-southern United States was mostly affected. California, particularly Los Angeles, consistently reported the highest case and death counts, likely influenced by its dense population and specific activity patterns. This underscores the need for targeted lockdown strategies in heavily impacted urban centers. Furthermore, the observed positive correlation between total cases and deaths, coupled with the identification of regions with exceptionally low death rates, indicates that while the overall mortality remained relatively low, proactive measures such as social distancing and targeted lockdowns in high-risk areas could have effectively curbed the pandemic's spread. These insights provide valuable guidance for future pandemic preparedness and response, emphasizing the importance of timely, data-driven interventions to mitigate disease transmission and mortality.