Understanding COVID-19 Dynamics

By Sayali Sali and James Field

Presentation Slides, Project Folder

Introduction:

The objective of our analysis is to gain a comprehensive understanding of the COVID-19 situation in the United States, with a specific focus on California and Texas. We aim to examine key metrics such as testing, positive cases, hospitalization, and deaths associated with the virus. By utilizing data from the COVID Tracking Project API, our goal is to offer insights into the spread of COVID-19 and assess the efficacy of testing and healthcare measures. This analysis holds significance for policymakers, healthcare professionals, and the wider community in making well-informed decisions regarding public health strategies and resource distribution

Dataset:

The dataset comprises daily COVID-19 statistics for each state in the USA, including total cases, recoveries, positive cases, deaths, hospitalizations, and testing details. The dataset is well-suited for analysis in the context of the COVID-19 pandemic as it provides granular insights into key metrics at the state level. Data obtained from the COVID Tracking Project API (https://covidtracking.com/data/api), extracted using Python's requests, saved as a JSON file, and processed into a DataFrame and then analyzed using Pandas and Matplotlib.

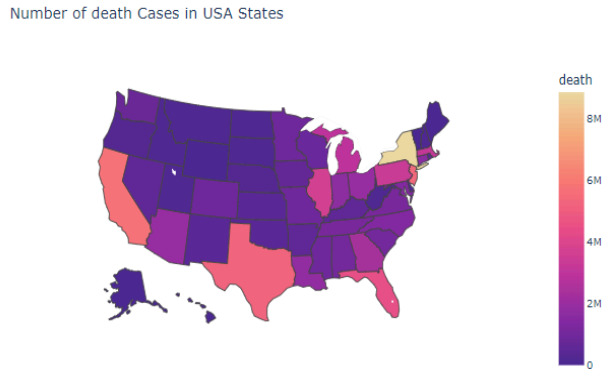
Analysis Technique:

We employed various analytical techniques, including data visualization, statistical testing, and correlation analysis. These techniques help us identify patterns, trends, and relationships within the data. Visualizations such as choropleth maps and time series plots offer intuitive representations of COVID-19 metrics, while statistical tests provide insights into the significance of observed differences. Correlation analysis helps uncover relationships between different variables, such as positive cases and deaths, aiding in understanding the underlying dynamics of the pandemic.

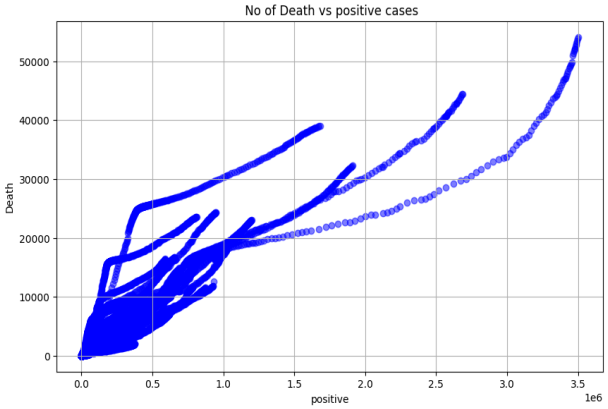
Results:

Our analysis revealed several key insights:

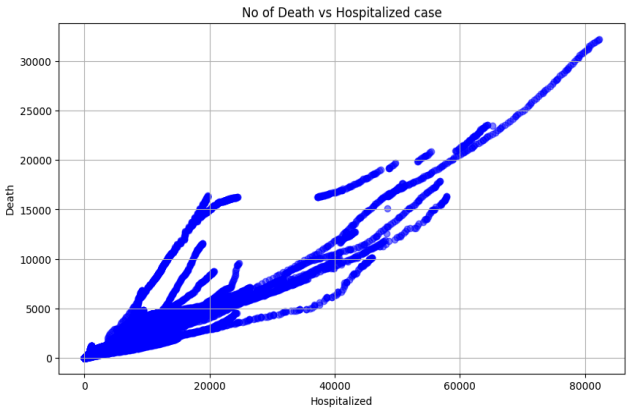
- The choropleth map illustrates the variation in COVID-19 death rates across different states, highlighting regions with higher mortality rates. Here we can see Texas and California having a high number of deaths.



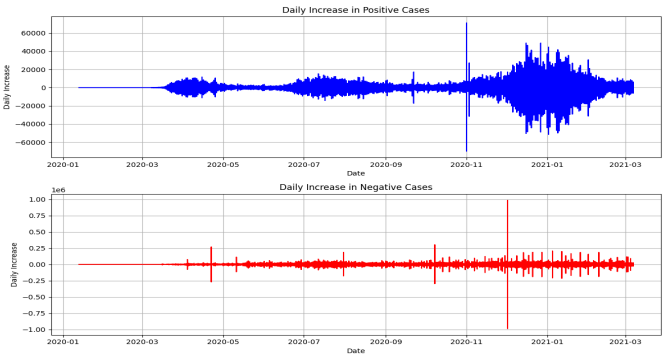
- To check if positive cases increase with increase in death rate we plot scatter plot and find correlation coefficient between the number of positive cases and deaths is 0.881, indicating a strong positive correlation. The p-value of 0.0 suggests that this correlation is statistically significant. This implies that as the number of positive cases increases, the number of deaths tends to increase as well.



- Next we looked into the number of hospitalized cases and deaths. The Pearson correlation coefficient between them is approximately 0.91, indicating a strong positive correlation. This suggests that as the number of hospitalizations increases, there is a corresponding increase in the number of deaths. The p-value of 0.0 indicates that this correlation is statistically significant, further supporting the observed relationship between hospitalizations and deaths. Thus, there is a clear association between the two variables, highlighting the critical role of hospitalization in predicting COVID-19-related fatalities.

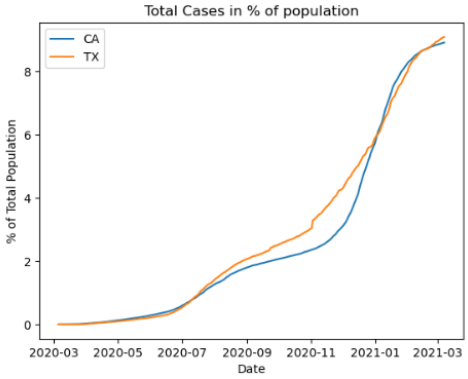


- To find the daily increase in positive and negative cases we calculated the difference in previous day and current day. We found that most positive cases

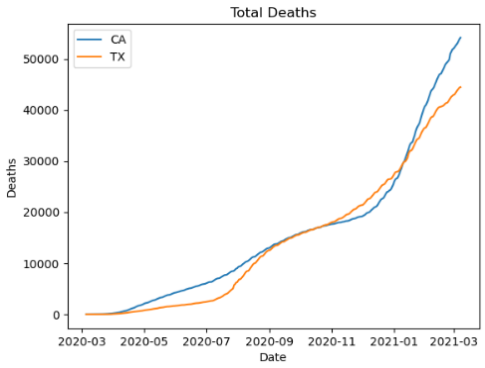


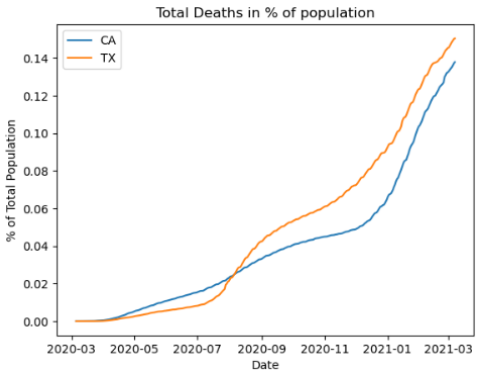
increased in the months of November 2020 to February 2021 and to cross check we found many news articles to confirm our analysis. But it was more interesting to see that negative cases were more detected than positive cases.

- We also checked the amount of covid cases that occurred in Texas and California and plotted them against each other. We then checked how many cases there were in regard to the population of each state showing us that the percent of cases when taking into account population were extremely similar.

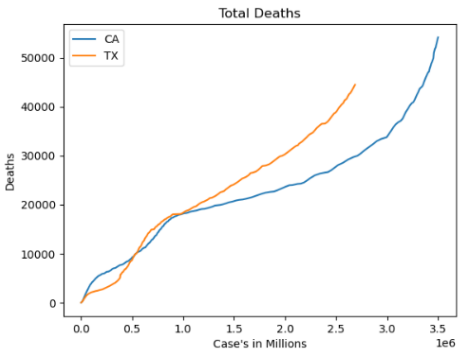


- Another test we conducted on our data was checking the total deaths in Texas and California. We first took the numbers of deaths in both states and then we again checked the amount of deaths as a percentage versus the population. We found that while California had more deaths, Texas had a higher percentage of their population who died from Covid by .0125% with California's death percent being 0.1379% and Texas' being 0.1505%.





- We also compared the number of cases to the number of deaths in these two states. From this we can easily see that while California had more deaths it is only because they had more confirmed cases. When California and Texas reported the same number of cases Texas had more deaths per case of Covid.



Technical:

Data Preparation: The dataset obtained from the API needed some preprocessing before analysis. This included converting date columns to datetime format, sorting the data by date, and converting certain columns to numeric data types. Additionally, missing values were handled appropriately to ensure accurate analysis results.

Analysis: The analysis conducted in this project is suitable for understanding the dynamics of COVID-19 cases and testing over time. Techniques such as correlation analysis, and visualization are common and effective methods for exploring relationships and trends in data. The use of these techniques allows for a comprehensive examination of how variables such as positive cases, deaths, and testing relate to each other.

Analysis Process: The analysis process involved several steps. Initially, the data was fetched from the API and preprocessed to ensure consistency and accuracy. Exploratory data analysis was then performed to understand the distributions and relationships between different variables. Various statistical tests, including correlation, were conducted to assess relationships and significance. Additionally, visualizations such as choropleth maps and time series plots were created to visualize trends over time. The process also involved experimentation with different parameters and techniques to refine the analysis and interpret the results effectively. Overall, the analysis process was iterative, involving multiple stages of exploration, testing, and refinement to gain meaningful insights from the data.