US COVID-19 Analysis

Objective

To study the total cases and deaths reported across the united states due to the COVID-19 pandemic in 2020 and 2021. In this way, the user can visualize as well as analyze the situation of covid across the country even at the county level. Moreover, it enables the user to compare the cases and deaths spike over a period between different counties in a state.

Dataset

The dataset used is from the NY Times GitHub(nytimes/covid-19-data: An ongoing repository of data on coronavirus cases and deaths in the U.S. (github.com) page which is also used by Google to give a summary on their home page. But unlike Google, the project enables the selection of datewise details of counties. The original dataset ranges from January 01, 2020, to Oct 11, 2021. But due to cloud space limitation, data is truncated from Oct 16, 2020.

In order to visualize the case distribution in the map, the dataset was concatenated with the latitude and longitude of each county from a public dataset.

Design Process

The dataset is very large enabling day-wise details for each county over a year. So it is very important the user is given choices for filtering features which are mainly the Date, State and County. Based on the there are different visualizations possible.

ScatterPlot Map- As per the date selected, the map highlights the coordinates of counties that reported the cases. Pydeck was another alternative but failed due to the large dataset.

Exact Cases and Details- A query box is created where the exact total number of cases and deaths reported as of that date, the state, and the county is implemented for detailed study.

Hot Spot Bar Graph- A bar graph is created highlighting the hot spot regions that reported most of the cases in the country. The user is allowed to add or delete the default selection. The same is replicated for the deaths reported. Area chart was another alternative that was not showing index as the county name.

Line Graph- A line graph showing the cases as well deaths reported enabling the user to see spikes over the entire time.

Development Summary

The entire application took around 50 hrs to develop. The major chunk was due to the delay in processing because of the size of the dataset. Moreover, the dataset had to be appended with the latitude and longitude of each county. Almost 60% of the time was only spent on the EDA part and around 10% was for setting up the environment and writeup.