Anveshrithaa Sundareswaran Andrew ID: asundare

05839-A: Interactive Data Science
Assignment 2

Objective:

To design and implement an interactive application using streamlit for **analyzing and visualizing COVID-19 statistics** including total number of cases, deaths, new cases, running average for every state and county in the United States to identify trends and insights from the data and answer key questions such as

1. Which are the states with the highest number of cases and mortality?
2. Are the number of cases and deaths proportional to each other for a given geographic location?
3. Which geographic location is more vulnerable and what is the trend in the number of cases and deaths?

based on user selection through dynamic query filters and interactive visualizations.

Motivation:

During the ongoing COVID-19 pandemic, it is very important to report accurate case numbers and other information about the outbreaks. The importance of this data cannot be understated as it can be used to gather useful insights and answer several key questions. This being the need of the hour, this application chooses to analyze and visualize COVID-19 data for gathering meaningful insights and identifying the most vulnerable demographics.

Dataset:

For this application, the **New York Times COVID-19 live dataset**[1] is used. This dataset is publicly available and frequently updated by the New York Times. It primarily consists of the cumulative counts of coronavirus cases and deaths in the United States, at the state and county level, over time. For this application, we use the county-level data from the counties.csv file (https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv)

Dynamic Query Filters:

- User can select the number of states that are to be displayed in bar chart
- User can choose the type of bar chart
  a) States with highest number of total COVID-19 cases
  b) States with highest number of deaths
- User can select the state and the county whose COVID-19 stats are to be displayed.
- User can choose the number of days for which the data will be displayed in the summary table
- User can choose the number of days taken into account for calculating the moving average of new cases in the selected county

Visualizations:

1. **Scatter plot for Total number of cases vs Total number of deaths for 10 states**
   - A scatter plot that visualizes the number of cases versus the number of deaths for the selected number of states.

- The x-axis represents the total cases and the y-axis represents the total deaths.
- Each state is marked with a different color which is described in the legend.
- The size of the marker of each state is proportional to the number of total cases. That is, the marker of a state with higher number of total cases is bigger than the size of the marker of a state with lower number of cases.
- When the user hovers over the points, information including the state name, cases and deaths is displayed (**tooltip**).
- This graph enables the user to identify the correlation between number of cases and deaths.
- This visualization can be used to **answer if mortality is proportional to the number of cases**. That is, if number of deaths is higher in a state with higher number of cases.
- From the plot, it can be observed that higher number of cases does not imply higher number of deaths. This suggests that the mortality is not proportional to the number of cases.
- For example, in the case of Pennsylvania and Illinois, even though the number of cases in Pennsylvania is lesser than that if Illinois, the number of deaths in PA is higher than Illinois.

2. **Bar chart for 'n' states with the highest number of cases and deaths**
   - This bar graph shows the total case count/death count for 'n' states with the highest number of cases/deaths based on the user selection.
   - The x-axis represents the states and the y-axis represents the number of cases/deaths.
   - This helps in easily **identifying the states that are most affected and least affected** by COVID-19.

3. **Summary table for the last 'n' days for the chosen county**
   - This table summarizes the COVID-19 statistics such as total cases, total deaths, new daily cases and running average for the last 'n' days for the county chosen by the user.
   - The number of new cases for a particular day for the chose county is calculated from the dataset by finding the difference between the total cases on the previous day and that day.
   - Similarly, the moving average of new cases is calculated by averaging the number of new cases over the past 'm' days.

4. **Line chart for total cases for the selected county**
   - This line chart plots the number of cases since the day the first case was reported in that county till present.
   - Line chart makes it easier to visualize the trend in the data that what a a tabular representation can do.
   - The chart can be panned and zoomed according to the user preference. Zooming in will let the user visualize it at a more granular level while zooming out will show the trend for a wider range of time.
   - When user hovers over the line chart, information such as the data and number of cases are displayed (**tooltip**).
   - This informs the user about **how rapid the increase in cases is.**

5. **Line chart for moving average for the selected county**
   - This line chart plots the moving average of the new daily cases for the past 'm' days, with similar properties as that of the previous chart.

- This **answers the question whether the number of COVID-19 cases is increasing, decreasing or constant in the selected county**.

Development process:

This assignment was done individually and it took around 10 hours to build the complete application and get it running. The process started with spending a significant amount of time in exploring the dataset and brainstorming ideas on how to visualize it to enable meaningful and actionable insights. This application was thoughtfully designed to provide a simple and user-friendly interface design with interactive elements. Next, a small amount of time was spent on getting familiar with Streamlit. The implementation of the application did not consume a lot of time as it was made easy by Streamlit. Most of the visualizations could be implemented with just one line of code which made it very simple.

References

1. https://github.com/nytimes/covid-19-data#geographic-exceptions
2. https://docs.streamlit.io/library/cheatsheet
3. https://github.com/ashaabrizvi/Covid-19_Dashboard
4. https://share.streamlit.io/panditpranav/svm_covid_tracking/main/COVID_app.py
5. https://share.streamlit.io/remingm/covid19-correlations-forecast/main.py