▾ Part B: Analyze the data

Rena Liu

Note: Corresponding pdf is a better formatted version of the graphs and responses, but code generating each graph can be found here (along with same copy of responses in blue).

```
# Configure Libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.dates as mdates

# Mount data from files on Google Colab
from google.colab import drive
drive.mount('/content/drive', force_remount=True)

# Pull data
data = pd.read_csv('LA_singlestate_timeseries.csv')
```

```
    Mounted at /content/drive
```

B.1: Data Analysis

1. Graph the relationship between actual covid cases and vaccinations completed. Are there any salient trends or strong relationships? For which time periods do these trends exist? Please share your key findings and attach your graph with covid cases as the x-axis. Your graph should be easy to understand.

```
# Create scatter plot of actual covid cases vs. vaccinations completed,
# with scatter size being small (size 2)
plt.scatter(data['actuals.cases'], data['actuals.vaccinationsCompleted'],
            s=2, label="Actual Covid Cases vs. Vaccinations Completed")

# Label axes
plt.xlabel("Actual Covid Cases")
plt.ylabel("Vaccinations Completed")

# Turn off scientific notation for population metrics
plt.ticklabel_format(style='plain')

# Add title, legend, and show graph
plt.title("Actual Covid Cases vs. Vaccinations Completed in Louisiana")
plt.legend()
```
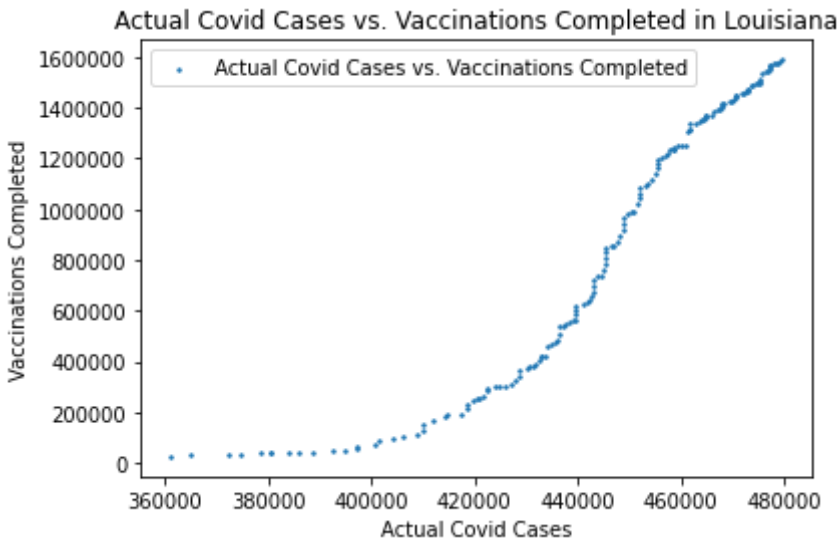
```
# Save graph as a png
plt.savefig('B1_Cases_vs_Vaccinations.png')

plt.show()
```



```
# Create scatter plot of vaccinations completed vs. actual covid cases,
# with scatter size being small (size 2)
plt.scatter(data['actuals.vaccinationsCompleted'], data['actuals.cases'],
            s=2, label="Vaccinations Completed vs. Actual Covid Cases")

# Label axes
plt.ylabel("Actual Covid Cases")
plt.xlabel("Vaccinations Completed (in Millions)")

# Add title, legend, and show graph
plt.title("Vaccinations Completed vs. Actual Covid Cases in Louisiana")
plt.legend()

# Save graph as a png
plt.savefig('B1_Vaccinations_vs_Cases.png')

plt.show()
```

Vaccinations Completed vs. Actual Covid Cases in Louisiana
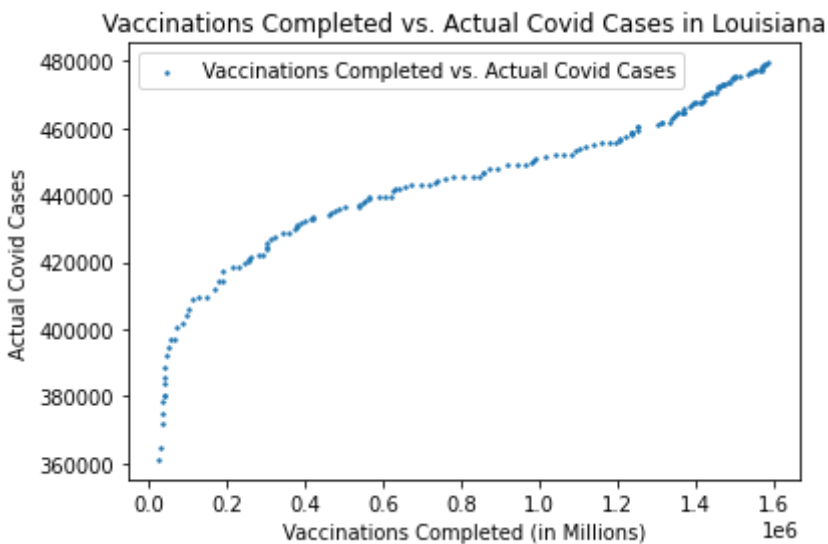
```python
# Create time series of vaccinations completed and actual covid cases over time
fig, ax = plt.subplots()
ax.plot(data['date'],data['actuals.cases'], label="Actual Cases")
ax.plot(data['date'], data['actuals.vaccinationsCompleted'],
        label="Vaccinations Completed")
plt.plot()

# Label axes
plt.xlabel("Date")
plt.ylabel("Population (in Millions)")

# Adjust date increments and labels using datetime library
plt.xticks(np.arange(0, len(data['date'])+1, 60),
           np.arange(datetime.strptime(data.loc[0]['date'], '%Y-%m-%d').date(), len(da

# Rotate and align the tick labels so they look better.
fig.autofmt_xdate()

# Add title, legend, and show graph
plt.title("Actual Covid Cases and Vaccinations Completed over Time in Louisiana")
plt.legend()

# Save graph as a png
plt.savefig('B1_Vaccinations_and_Cases_over_Time.png')

plt.show()
```
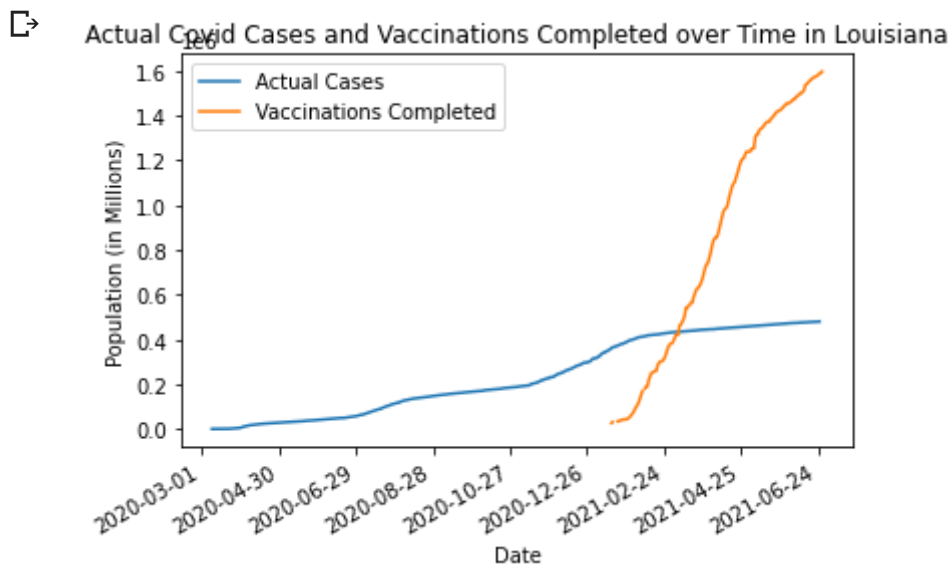
Actual Covid Cases and Vaccinations Completed over Time in Louisiana

There are three key trends seen in the relationship of actual covid cases and vaccinations completed in Louisiana; first during the period of 0 to around 420,000 covid cases, where there is a dramatic increase of cases, but little to no growth in vaccinations completed (Fig 1). Next, as the 400,000 covid cases grows to around 460,000 cases, the number of vaccinations also grows exponentially. Finally, as cases grows from 460,000 to 480,000, vaccinations completed grows at a strong positive linear relationship.

The impact of vaccinations completed on covid cases, which may be more relevant to our users, can be better reflected in Fig 2 and 3, where we graph vaccinations completed on the x-axis and a timeseries of both vaccinations completed and covid cases. Here, we see a period where cases grow dramatically but there are little to no vaccinations. Then, there is a period where case growth slows down as vaccinations increase exponentially. Finally, there is a period where cases and vaccinations both increase at a more constant rate.

2. Graph a time series of new covid deaths. Are there any salient trends you can see? For which time periods do these trends exist? Please share your key findings and attach your graph with covid cases as the x-axis. Your graph should be easy to understand.

```
# Create time series of actual covid deaths over time
fig, ax = plt.subplots()
plt.plot(data['date'], data['actuals.newDeaths'], label="New Covid Deaths")
plt.plot()

# Label axes
plt.xlabel("Date")
plt.ylabel("New Covid Deaths")

# Adjust date increments and labels using datetime library
```
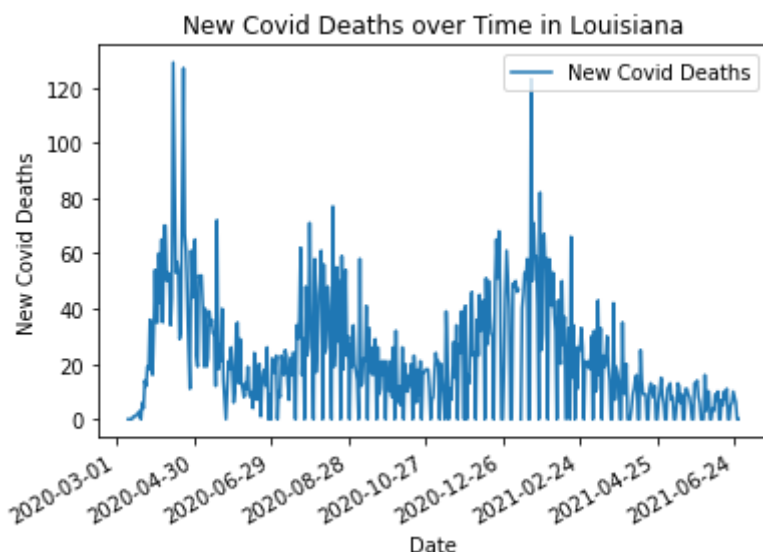
```
plt.xticks(np.arange(0, len(data['date'])+1, 60),
           np.arange(datetime.strptime(data.loc[0]['date'],
                                       '%Y-%m-%d').date(), len(data['date'])+1, 60))
# Rotate and align the tick labels so they look better.
fig.autofmt_xdate()

# Add title, legend, and show graph
plt.title("New Covid Deaths over Time in Louisiana")
plt.legend()

# Save graph as a png
plt.savefig('B1_Deaths_over_Time.png')

plt.show()
```


As time proceeds, there are three peaks of high numbers of new covid deaths, corresponding to three waves of time periods when new covid deaths grew dramatically in Louisiana (Fig 4). These three peaks occur between time periods March 2020 to June 2020, June 2020 to October 2020, and October 2020 to April 2021.

B.2: Open-Ended Questions

Note: These questions are not necessarily related to data analysis.

3. Why do you think there is very limited data on hospital and ICU bed usage in Pennsylvania before mid-May 2020? Limit your answer to 1-2 sentences.

My initial hypothesis explanation for why there is limited data is due to the historical lack of protocol for counting, tracking, and reporting bed usage numbers. Even if protocols existed for decisions like who is responsible for tracking and how to most effectively keep track of the quickly changing status of bed usage numbers, the extreme chaos of overwhelmed staff and resources likely lowered the priority of data collection for the sake of national level data sharing compared to urgent issues like keeping patients and staff (both hospital and governmental agency collecting and aggregating the data) safe, sane, and healthy during the pandemic.

4. Given what you know about covid, refer to your time-series graph to explain which societal factors, if any, may have impacted the number of new covid deaths. Please hypothesize why you believe so. State any assumptions you have made. Limit your answer to no more than 2 paragraphs.

Based off my best understanding, the factors impacting the various growths and declines of new covid deaths can be categorized into two key societal factors: 1. individual behaviors that impact the infection and transmission rates, and 2. broader access and institutional quality of medical care. The general public's increased uptake in behaviors that decrease the infection and transmission rates results in fewer new covid deaths, as does better access and quality of care.

Specific examples within the first category of individual actions include incentives for society to both stay informed about and follow public health recommendations: getting tested, wearing masks, getting vaccinated, frequent handwashing, and social distancing and isolation. The rate of uptake of these behaviors in turn are also impacted by factors like perception of risk weighed against other personal interests: trust in the government and medical institutions, socioeconomic stability and need to keep working, and personal interests for the activities sacrificed in participating in those public health recommendations in like maintaining one's social networks. Beyond individual actions, the category of institutional quality includes widespread and equitable access to the resources that enable the uptake of the aforementioned behaviors: testing and vaccination facilities, hospital resources and capacity (beds, staff, ventilators, and PPE, among others), and socioeconomic support for social isolation (for example housing support and technological and structural support for decentralized work, education, and childcare). As explained above, I hypothesize that the promotion of these positive factors reduced the spread and therefore deaths from covid and the decline of support for these factors increased the spread and deaths.

5. What are some possible limitations of this dataset? Limit your answer to 2-3 bullet points.

Since our key purpose of the dataset is to support the state-official users planning the re-opening of offices and business services in the US, it would be helpful for them to be informed of which localities are doing well and what societal factors and policies are correlated with these localities doing well. Specific limitations the dataset related to this are:

- Lack of breakdown of currently existing fields (like cases, deaths, ICU beds) by state county
- Lack of any quantifiable societal and policy factors (enforcement of a stay-at-home order, mask guidelines, permission for outdoor, indoor, or neither business)
- Lack of quality assurance of data accuracy and missing values

6. Briefly explain 1-2 recommendations you would relay to the team that maintains this dataset to overcome any of the limitations you mentioned above. Please keep your recommendations related to improving data quality, and not improving Louisiana's overall covid response. Show

your abilities to explain and persuade through writing. Limit your answer to no more than 2-3 sentences.

To address the limitations brought up above, I recommend that the dataset team maintain close communication with data providers (hospitals and other medical and governmental institutions) to obtain increased data preciseness (which county each hospital is reporting from) as well as the auxiliary pieces of data (regarding the societal and policy factors of each county at each date). This close communication will also ensure transparency and collaboration when it comes to identifying reducing data errors and missing data.