Schmidt Technical Assessment - Jacob Rubin

Part A

In [184]:

```
import pandas as pd
import datetime
import sys
```

In [185]:

```
def dashboard(date):
    # Create the dataframe
    dupDf = pd.read_csv("PA_singlestate_timeseries.csv")

    # Find number of duplicate rows
    dups = len(dupDf)-len(dupDf.drop_duplicates())

    # Use dataframe without duplicates
    df = dupDf.drop_duplicates()

    # Access data row for given date
    juneData = df.loc[(df["date"]=="2021-06-16")]

    # Display date as m/d/y instead of year-month-day
    displayDate = datetime.datetime.strptime(date, '%Y-%m-%d').strftime('%m

    # Getting the data we need from the dataframe
    newCases = int(juneData['actuals.newCases'].values[0])
    newDeaths = int(juneData['actuals.newDeaths'].values[0])
    cumCases = int(juneData['actuals.cases'].values[0])
    cumDeaths = int(juneData['actuals.deaths'].values[0])
    casesPer100KPop7DayAvg = juneData['metrics.caseDensity'].values[0]
    actualCompVacc = int(juneData['actuals.vaccinationsCompleted'].values[0
    ratioCompVacc = juneData['metrics.vaccinationsInitiatedRatio'].values[0
    actualInitVacc = int(juneData['actuals.vaccinationsInitiated'].values[0
    ratioInitVacc = juneData['metrics.vaccinationsCompletedRatio'].values[0

    # Dashboard to print
    sys.stdout.write('\n\033[1mDataset Overview: \033[0m\nThis dataset incl
    sys.stdout.write("\033[1mPennsylvania, " + str(displayDate) + '\033[0m\
    sys.stdout.write("\t - New Cases: " + str(newCases) + '\n')
    sys.stdout.write("\t - New Deaths: " + str(newDeaths) + '\n')
    sys.stdout.write("\t - Cumulative Cases: " + str(cumCases) + '\n')
    sys.stdout.write("\t - Cumulative Deaths: " + str(cumDeaths) + '\n')
    sys.stdout.write("\t - Cases per 100k (using a 7-day rolling average):
    sys.stdout.write("\t - Completed: " + str(ratioCompVacc) + '  (' + str(
    sys.stdout.write("\t - Initiated: " + str(ratioInitVacc) + '  (' + str(

    return ""

if __name__ == '__main__':
    dashboard("2021-06-16")
```

```
Dataset Overview:
This dataset includes Pennsylvania COVID-19 data for the time period from
3/1/20 to 6/28/21. Data includes total/daily cases, deaths, and vaccinati
ons, as well as hospital and ICU capacity. Data is presented in both tota
l numbers and metric ratios. There are 20 duplicate rows throughout the d
ataset.

Pennsylvania, 06/16/21

Infection:
         - New Cases: 279
         - New Deaths: 16
         - Cumulative Cases: 1214051
         - Cumulative Deaths: 27582
         - Cases per 100k (using a 7-day rolling average): 2.9

Vaccination:
         - Completed: 0.611  (6045055)
         - Initiated: 0.472  (7818311)
```

Part B

In [186]:

```
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
```

In [187]:

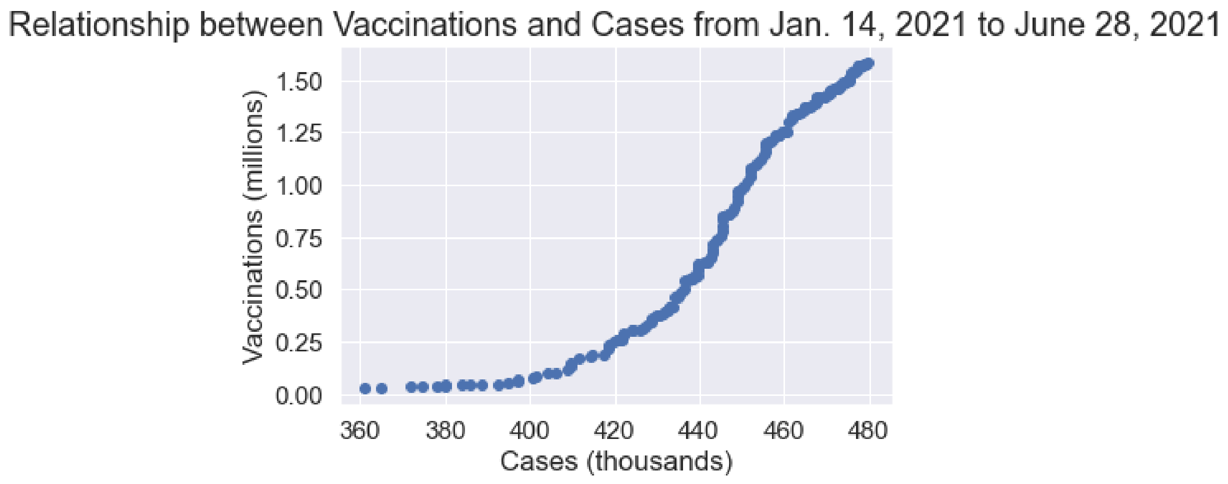
```
# Create the dataframe
df = pd.read_csv("LA_singlestate_timeseries.csv")
#df.head()

dates = df['date'].values
actualCompVacc = df['actuals.vaccinationsCompleted'].fillna(0).values
actualCompVaccInMils = (df['actuals.vaccinationsCompleted'].values / 100000
newCases = df['actuals.newCases'].values
newDeaths = df['actuals.newDeaths'].values
cumCases = df['actuals.cases'].values
cumCasesinThds = (df['actuals.cases'].values / 1000.0)
newDeaths = df['actuals.newDeaths'].values

jan = (df['actuals.cases'][df.date=='2021-01-14'].values / 1000.0)
feb = (df['actuals.cases'][df.date=='2021-02-14'].values / 1000.0)
mar = (df['actuals.cases'][df.date=='2021-03-14'].values / 1000.0)
apr = (df['actuals.cases'][df.date=='2021-04-14'].values / 1000.0)
may = (df['actuals.cases'][df.date=='2021-05-14'].values / 1000.0)
june = (df['actuals.cases'][df.date=='2021-06-14'].values / 1000.0)


#B.1.1 actual covid cases vs. vaccines completed
plt.scatter(cumCasesinThds, actualCompVaccInMils)
sns.set(font_scale=1.3)
plt.xlabel("Cases (thousands)")
plt.ylabel("Vaccinations (millions)")
plt.title("Relationship between Vaccinations and Cases from Jan. 14, 2021 t

plt.show()
```


Relationship between Vaccinations and Cases from Jan. 14, 2021 to June 28, 2021

Are there any salient trends or strong relationships? For which time periods do these trends exist?

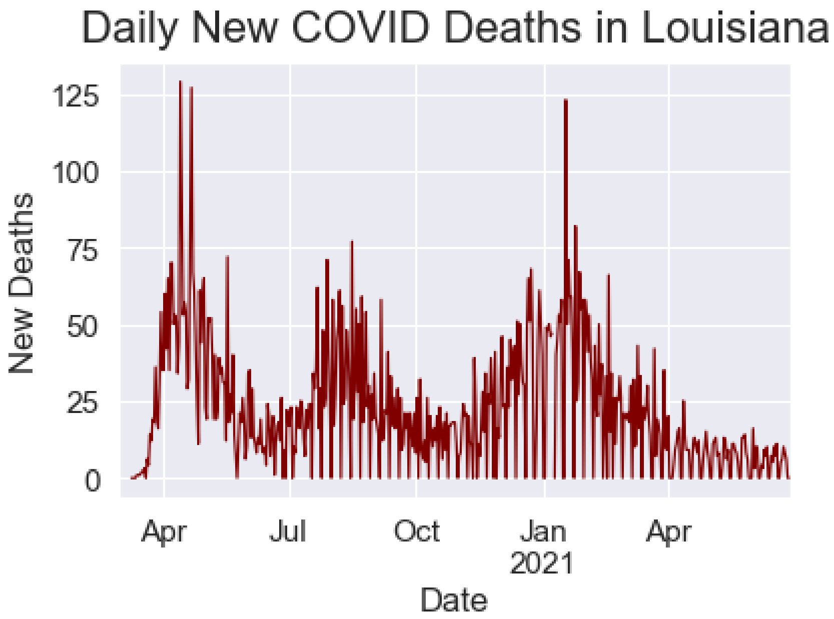
The graph demonstrates a direct relationship between cases and vaccines from the time period from Jan. 2021 through June 2021. While we cannot determine that an increase in cases directly causes an increase in vaccinations (or vice versa), we note the positive correlation between the two.

In [190]:

```
# B.1.2

# Creating time series plot
df_cases = pd.DataFrame({'date': dates, 'newDeath': newDeaths})
df_cases['date'] = pd.to_datetime(df_cases['date'])

# Styling Plot
sns.set(font_scale=1.4)
df_cases.set_index('date')['newDeath'].plot(figsize=(6, 4), linewidth=1.2,
plt.xlabel("Date")
plt.ylabel("New Deaths")
plt.title("Daily New COVID Deaths in Louisiana", y=1.02, fontsize=22);
```



Are there any salient trends or strong relationships? For which time periods do these trends exist?

Reviewing the graph, we note three significant spikes in deaths during the following time periods: (1. May-June 2020, 2. Aug.-Sept. 2020, 3. Jan-Feb. 2021). In this time-series, we note that spikes in deaths occur in approximately two-month waves, indicating periods of rise and decline. These spikes may be associated with the initial rise of the pandemic, a post-lockdown rise, and holiday spreader events during winter 2021.

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.

Nuances in Pennsylvania's definitions of hospitalization and ICU admittance and a decentralized reporting system hampers data collection. Additionally, state officials must make difficult decisions to negotiate patient privacy/transparancy and focus limited resources on improving contact tracing and COVID testing rather than data reporting.

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

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.

I hypothesize that the factors of (1) political polarization, (2) economic inequity, and (3) social vulnerability disproportionately impact new COVID-19 deaths. Reviewing the time-series graph, we note three distinct spikes in COVID-19 deaths. The first spike (May-June 2020) we attribute to the initial rise of the pandemic. During this time period, political polarization encouraged the downplaying of the spread and magnitude of the virus. In turn, misinformed people did not take proper safety precautions such as social distancing and wearing masks. Additionally, inequities in poverty and inadequate access to health care caused underserved communities to suffer. Those without proper treatment had an increased risk of getting COVID, harsh symptoms, and death. Further, social vulnerability is key in these trends. Older generations, immunocompromised individuals, and densely populated communities are more likely to contract COVID. The greatest clusters of COVID-19 often occur in these environments such as nursing homes and prisons. In effect, COVID-19 ravaged those who could not properly defend themselves. These trends continue with the second and third spikes. Looking at the second post-lockdown spike (Aug-Sept 2020) and third winter holiday spike (Jan-Feb 2021), the same three factors manifested themselves to cause deaths. Political polarization allowed maskless individuals to pass the virus at super-spreader events, social inequities cause under-resourced communities to remain underserved, and socially vulnerable individuals are disproportionately affected. This hypothesis is based on the assumptions that the LA time-series graph is reflective of the greater United States.

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

- Scale. Finer-grain data on the level of counties, precincts, or rural/urban areas may better track the movement and clustering of COVID-19 cases.
- Patient data. Aggregated demographic patient data with respect to age, ethnicity, and household income may better track the societal underpinnings of COVID-19 cases.
- Context. Comparative metrics of Pennsylvania against neighboring states, regions, and the entire United States may better track the welfare of the state and its population.

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. We are looking for how well you can explain and persuade through writing. Limit your answer to no more than 2-3 sentences.

1. The dataset should be delineated at the county/precinct level, in addition to the current state level.
2. The dataset should be presented in comparison to other states, regional data, and the country as a whole.

   These recommendations expand the scope of the data without extra data collection, providing scale and context that encourages deeper social and health analysis.

```
In [ ]:
```