towards
data science

Follow 594K Followers

Visualizing NYS COVID-19 Data

With Python Matplotlib and Plotly

Meagvo 1 day ago · 5 min read★

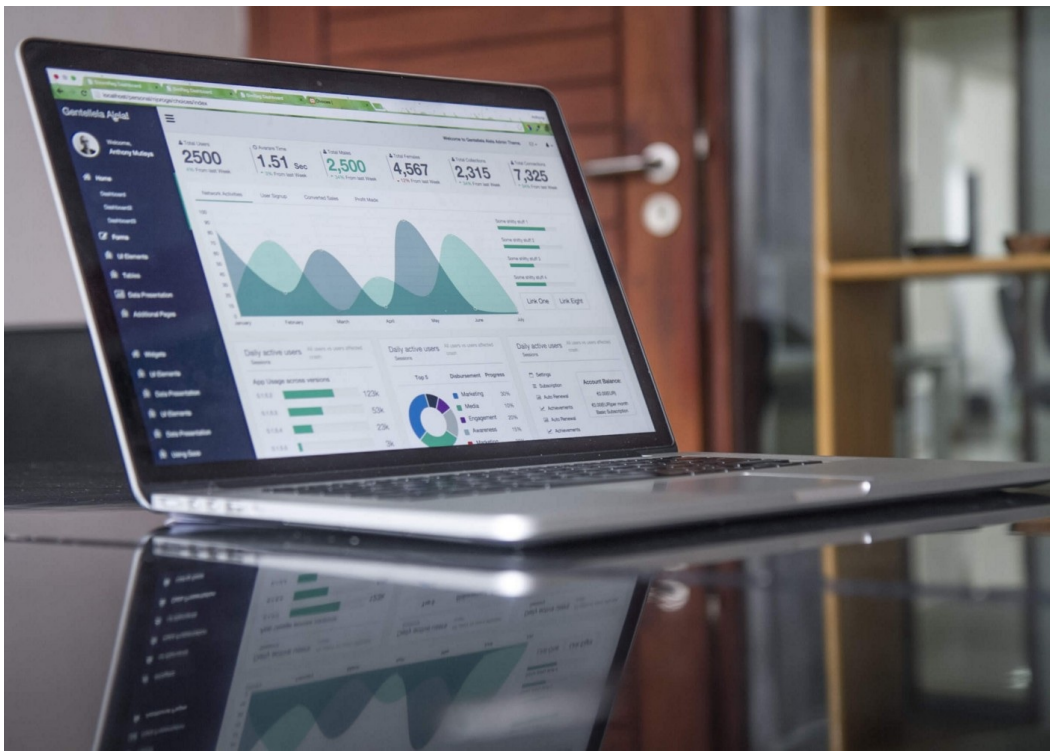

Photo by Carlos Muza on Unsplash

. . .

As a full-time Data Analyst, I spend much of my day using SQL to extract the relevant data for my work projects. However, on the side, I also like to hone my python skills with exploratory research and graphics. The following (simple) project focuses on using plotly to graph out the New York State COVID-19 data from Health.Data.NY.Gov.

The Goal

The goal of this project was to start with a simple visualization created from the matplot library and then enhance it and build upon it using plotly.

The Data

I used two datasets for my descriptive analysis and visualizations:

1. COVID-19 Testing Data
2. COVID-19 Vaccination Data

Both datasets contain up-to-date statistics from NYS by county.

After I imported the data, I did a quick summary to understand the structure. First, I looked at the COVID testing data:

Test Data.describe

A quick describe function showed me the basics. My data frame has 37882 rows and 6 columns. The columns that I will be focusing on are County, Test Date, and Cumulative Number of Tests Performed.

. . .

```
<bound method NDFrame.describe of                 Region       County  First Dose  Series Complete Report as of
0        Capital Region       Albany          69               0   12/14/2020
1      Western New York     Allegany           0               0   12/14/2020
2         New York City        Bronx           2               0   12/14/2020
3          Southern Tier       Broome           0               0   12/14/2020
4      Western New York  Cattaraugus           0               0   12/14/2020
...                 ...          ...         ...             ...          ...
19835    Capital Region   Washington       37189           35263   10/29/2021
19836      Finger Lakes        Wayne       57312           53892   10/29/2021
19837        Mid-Hudson  Westchester      750587          675291   10/29/2021
19838      Finger Lakes      Wyoming       20755           19377   10/29/2021
19839      Finger Lakes        Yates       12951           11936   10/29/2021

[19840 rows x 5 columns]>
```

Vaccine Data.describe

For the vaccine data, quick describe function showed me the basics as well. There are 1980 rows and 5columns. The columns that I will be focusing on are Region, County, Report as of, and First Dose.

Matplotlib Visual

My first visual was created from matplotlib using the overall data for the testing dataset. I used the following block of code (which I do not recommend if you want a meaningful visualization).

```
plt.plot(df['Test Date'],df['Cumulative Number of Tests Performed'])
plt.title('COVID Testing NYS')
plt.xlabel('Test Date')
plt.ylabel('Cumulative Count of Tests')
```

```
plt.show()
```


As you can see, this graph is not telling a good story, mainly because all of the county data is being plotted together. In addition, you can't see the test dates because there are far too many points. Therefore, the first steps I took were to clean the data. I focused my analysis on Long Island, and therefore I selected Suffolk County and Nassau County.

First, I set the datetime index, grouped by count, and resampled:

```
dftest=df

dftest['date']=pd.to_datetime(df['Test Date'])

dftest=df.set_index('date')

dfcounty1=dftest.groupby(['County']).resample('D').mean().reset_index
()
```

Next, I separated the Suffolk County data from the Nassau County data:

```
Suffolk=dfcounty1[dfcounty1['County']=='Suffolk']
Nassau=dfcounty1[dfcounty1['County']=='Nassau']
```


This is a little better, but it's still messy, and it's not giving me all the information I want. By this point, I decided that my goal would not be accomplished with matplotlib, so I moved onto plotly.

Basic Plotly Visual

This simple block of code allowed me to plot out the Suffolk County data:

```
fig = px.line(dfcounty2, x = 'date',
y='Cumulative Number of Tests Performed', color='County',
title = 'COVID Tests by County')
fig.show()
```

Now, this was more of the format I was looking for, but now it was time to combine my data. The code below allowed me to plot the testing data and the vaccine data together for Nassau and Suffolk County.

To break it down, first I created subplots to allow for a dual y-axis. Then I plotted the Vaccine data for Suffolk and Nassau. I added traces for each line that I wanted to add. Then I formatted the titles and added annotations. I wanted to highlight the date of the first COVID test on LI and the first COVID vaccine on LI, so I added that with the text annotations.

Finally, I created my dropdown filter with the updatemenus parameter.

```
fig2 = make_subplots(specs=[[{"secondary_y": True}]])
fig2.add_trace(go.Scatter(x=S['date'], y=S['First Dose'], name =
'Suffolk First Dose of Vaccine',
                          line=dict(color='red', width=1,
dash='dash')), secondary_y=True)
fig2.add_trace(go.Scatter(x=N['date'], y=N['First Dose'], name =
'Nassau First Dose of Vaccine',
                          line=dict(color='blue', width=1,
dash='dash')), secondary_y=True)

fig2.add_trace(go.Scatter(x=Suffolk['date'], y=Suffolk['Cumulative
Number of Tests Performed'], name = 'Suffolk Tests',
                          line=dict(color='red', width=1,
dash='solid')), secondary_y=False)
fig2.add_trace(go.Scatter(x=Nassau['date'], y=Nassau['Cumulative
Number of Tests Performed'], name = 'Nassau Tests',
                          line=dict(color='blue', width=1,
dash='solid')),secondary_y=False)
```

```
fig2.update_layout(title_text='COVID Test and Vaccines LI',
title_x=0.4,
                   xaxis_title='Date')
fig2.update_yaxes(title_text="COVID Tests", secondary_y=False)
fig2.update_yaxes(title_text="First Dose of Vaccine (Count)",
secondary_y=True)

fig2.add_annotation(x='2020-03-05', y=100,
                      xref="x",
        yref="y",
        text="First Test: 3/5/20",
        showarrow=True,
        font=dict(
            family="Courier New, monospace",
            size=10,
            color="black"
            ),
        align="center",

        ax=20,
        ay=-30,
        bordercolor="black",
        borderwidth=1,
        borderpad=2,
        bgcolor="#ffffff",
        opacity=1
        )
fig2.add_annotation(x='2020-12-14', y=20,
                      xref="x",
        yref="y",
        text="First Vaccine: 12/14/20",
        showarrow=True,
        font=dict(
            family="Courier New, monospace",
            size=10,
            color="black"
            ),
        align="center",

        ax=20,
        ay=-30,
        bordercolor="black",
        borderwidth=1,
        borderpad=2,
        bgcolor="#ffffff",
        opacity=1
        )

fig2.update_layout(

    updatemenus=[
        dict(active=0,
            buttons=list([
            dict(label="All",
                 method="update",
                 args=[{"visible":[True,True,True, True]},
                       {"title":"ALL"}]),
            dict(label="Suffolk",
                 method="update",
                 args=[{"visible":[True, False, True, False]},
                       {"title":"Suffolk"}]),
            dict(label="Nassau",
                 method="update",
                 args=[{"visible":[False,True,False, True]},
                       {"title":"Nassau"}]),
            dict(label="First Dose",
                 method="update",
                 args=[{"visible":[True, True, False,False]},
                       {"title":"First Dose"}]),

            dict(label="Total Tests",
                 method="update",
                 args=[{"visible":[False,False, True, True]},
                       {"title":"Total Tests"}])

        ]),
        )
    ]
```

```
)

fig2.show()
```

Findings

Overall, I found that Nassau county and Suffolk county followed very similar trends in terms of the COVID testing and vaccines. However, there were a few differences:

1. The first test was in Nassau county and it was positive. Therefore, I couldn’t accurately plot the positive rate at the beginning of the pandemic because Nassau would show up as 100% positive for the first day of testing and that makes sense because people weren’t being tested unless they were fairly certain they were positive.
2. Despite having more tests in Suffolk county, Nassau county was quicker to get the first dose of the vaccine. This also makes sense because Suffolk county is (generally) more conservative/republican and Nassau is (generally) more liberal as it is closer to NYC.

Thank you for taking the time to read this article. I understand that it is fairly basic, but I am hoping this acts as a good introduction for those who would like to use plotly to visualize data.

Please feel free to leave comments related to the findings or visuals and I will be sure to update them with new data.

Interactive public plotly figure here: https://chart-studio.plotly.com/~meagvo/2/#/

Full notebook available on github: https://github.com/meagvo/TowardsDataScience/blob/main/HealthProject.ipynb

Plotly Python Matplotlib Free Python Tutorials