Analysis of the CDC's data on Covid19 from January 2020 to January 2022

Romzy Safadi

March 10th, 2022

Introduction

Overview of Project and specific goals:

The overall goal of this product was to see if I could accurately show covid-19 trends throughout the pandemic with the data that the CDC provided. More specifically, a deeper dive into which states were effected the most, as well as how each varient effected the overall population within the United States. I wanted to accomplish this through accurate visual representations of the CDC's data. The data was grouped by state, from January 20th, 2020, all the way to January 28th, 2022. The main parts that I kept from the original data frame included total cases, number of new cases daily, total deaths, new deaths daily, submission dates, and the main 50 states + American territories.

Explain why this project was import to you

This proect was important to me as this was something that not only effected my life, but the lives of billions across the World. It's effected both my personal and professional career. Since the pandemic, our lives have been changed. We all had to adapt to working remotely, learning remotely, being careful as to who we see, when we see them, and how we see them. For many of us, this has played a role in our lives to this day, it could even effect our life style for the rest of our lvies. Some of us may never return to fully working in person. For me personally, I was not able to see my mother for nearly two whole years because of the pandemic. She is considered a high risk individual, due to prior health problems. Being that I lived not more than an hour away from her, this was unfortunate. My undergraduate program completely switched to online, I was not able to attend any labs, and even had a very unique graduation ceremony. My going out habits, and views on the world have also been effected during the pandemic. When the pandemic first started, I was constantly keeping up with news outlsets to find the latest news on COVID-19 and how it was effecting our day to day throughout the country. Eventually, after being on lockdown for so long, I just got used to staying indoors and doing my part. I no longer kept up with the trends and current events of COVID-19. This is why I wanted to take a look at the data, and map it out for myself. To see if I could take a look at the pandemics story, as a whole over the past two years. I wanted to see if I could identify when certain variants were present, how they effected us as a country, and other trends throughout.

Explain your hypothesis (or hypotheses) you set out to test. What hunch(es) did you have about the data?

My first hypothesis was that states with more leniant mask mandate like Florida and Texas, would have the highest number of cases, deaths, and suffered the most throughout the pandemic (ratio of deaths per cases). The second hypothesis that I focused on was that I would be able to clearly see when the original first few variants, delta varient, and Omicron varient were present within the United States. A hunch I had when going through the data was that California and New York would also be largely suffering from the pandemic as far as number of cases, deaths, and deaths per cases.

Provide a link to your data source if applicable

I found my data on the CDC's website:

https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36

The data can be exported in the top right, and viewed at the bottom of the page.

A brief overview of the steps needed to clean the data

Cleaning the data was not as straight forward as I thought it would be. In fact, I found myself working to clean the data through multiple weeks in order to obtain what I needed for each EDA phase.

For starters the original data frame had a few columns that I wanted to drop such as: conf_cases, prob_cases, pnew_case, conf_death, prob_death, pnew_death, created_at, consent_cases, and consent_deaths.

This left me with five columns: submission_date state, tot_cases, new_case, tot_death, and new_death.

I then had to remove all commas in my df so that I could group my data and organize it as I would like.

This was followed up by converting my df objects to integers so that I could better work with them.

Next, since the state column was in abbreviations, such as 'CA' instead of California, I converted all of the states to full names, as I thought this would be more visually appealing while plotting and reading the data.

Throughout the EDA process I found that the American territories were serving as major outliers when plotting, so I decided it would be best to drop them in order to keep the plots and project focused on the main 50 states.

I also created a data frame that was grouped by state on the last day of my recoreded data, January 8th, 2022. This was used to plot all of the total values from my data.

Lastly, admitedly one I struggled with a lot, was that the CDC placed NYC as a seperate input to NY state. This was very confusing and I eventually had to combine NYC and NY to form one state as it was messing with my plotting results.

I eventually added a new column called, percentage_of_Deaths_by_cases, which was the ratio of the total deaths per total cases for each state. This was used to identify help identify which states suffered the most.

Data Dictionary

Submission_date: Date when data was submitted to the CDC, should be daily by each 50 states.

state: The name of the state

tot_cases: Accumulative number of cases for that state on that day + all the days previous in the dataframe.

new_case: The accumulative number of new COVID-19 cases on that submission date for that specific state.

tot_death: Accumulative number of deaths for that state on that day + all the days previous in the dataframe.

new_death: Total number of deaths on that submission date for that specific state.

percentage_of_deaths_by_cases: This is the ratio of total deaths per total cases on the final submission date of January 28th, 2022.

Results

Read in final data

In [27]:
```
import pandas as pd
import numpy as np
import seaborn
import matplotlib.pyplot as plt
import plotly.express as px


#multiple outputs per cell

from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
```

In [5]:
```
#final data file:

df = pd.read_csv('Romzy_Safadi_covid19.csv')
df.info()
df.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 44280 entries, 0 to 44279
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   submission_date  44280 non-null  object
 1   state            44280 non-null  object
 2   tot_cases        44280 non-null  object
 3   conf_cases       23957 non-null  object
 4   prob_cases       23885 non-null  object
 5   new_case         44280 non-null  object
 6   pnew_case        40408 non-null  object
 7   tot_death        44280 non-null  object
 8   conf_death       23675 non-null  object
 9   prob_death       23675 non-null  object
 10  new_death        44280 non-null  object
 11  pnew_death       40305 non-null  object
 12  created_at       44280 non-null  object
 13  consent_cases    36895 non-null  object
 14  consent_deaths   37638 non-null  object
dtypes: object(15)
memory usage: 5.1+ MB
```

Out[5]:

	submission_date	state	tot_cases	conf_cases	prob_cases	new_case	pnew_case	tot_dea
0	1/28/22	GA	2,346,518	1,824,347	522,171	18,785	2,900	32,8(
1	1/28/22	PR	453,669	257,022	196,647	3,404	2,101	3,8(
2	1/28/22	FL	5,501,599	NaN	NaN	22,705	5,753	64,6
3	1/28/22	UT	875,251	875,251	0	6,166	0	4,1
4	1/28/22	HI	208,253	NaN	NaN	1,848	207	1,1

These are the main data frames I used throughout the project:

df3 is when all states were put into full names, and all columns that needed to be dropped were dropped:

In [12]:

```
#dropping columns to create df3

df3 = df.drop( columns = ['conf_cases', 'prob_cases', 'pnew_case', 'conf_deat
          'created_at', 'consent_cases', 'consent_deaths'])

#removing commas in df
```

```
df3['tot_death'] = df3['tot_death'].str.replace(',','')
df3['tot_cases'] = df3['tot_cases'].str.replace(',','')
df3['new_case'] = df3['new_case'].str.replace(',','')
df3['new_death'] = df3['new_death'].str.replace(',','')


#convert obects to int
df3["tot_cases"] = df3["tot_cases"].apply(pd.to_numeric)
df3["new_case"] = df3["new_case"].apply(pd.to_numeric)
df3["tot_death"] = df3["tot_death"].apply(pd.to_numeric)
df3["new_death"] = df3["new_death"].apply(pd.to_numeric)

# converting State abreviations to full name (ex: CA --> California)


df3["state"].replace({    "AL":"Alabama",
                      "AK":"Alaska",
                      "AZ":"Arizona",
                      "AR":"Arkansas",
                      "CA":"California",
                      "CO":"Colorado",
                      "CT":"Connecticut",
                      "DC":"Washington DC",
                      "DE":"Delaware",
                      "FL":"Florida",
                      "GA":"Georgia",
                      "HI":"Hawaii",
                      "ID":"Idaho",
                      "IL":"Illinois",
                      "IN":"Indiana",
                      "IA":"Iowa",
                      "KS":"Kansas",
                      "KY":"Kentucky",
                      "LA":"Louisiana",
                      "ME":"Maine",
                      "MD":"Maryland",
                      "MA":"Massachusetts",
                      "MI":"Michigan",
                      "MN":"Minnesota",
                      "MS":"Mississippi",
                      "MO":"Missouri",
                      "MT":"Montana",
                      "NE":"Nebraska",
                      "NV":"Nevada",
                      "NH":"New Hampshire",
                      "NJ":"New Jersey",
                      "NM":"New Mexico",
                      "NYC":"New York",
                      "NY" : "New York",
                      "NC":"North Carolina",
                      "ND":"North Dakota",
                      "OH":"Ohio",
                      "OK":"Oklahoma",
```

```
                            "OR":"Oregon",
                            "PA":"Pennsylvania",
                            "RI":"Rhode Island",
                            "SC":"South Carolina",
                            "SD":"South Dakota",
                            "TN":"Tennessee",
                            "TX":"Texas",
                            "UT":"Utah",
                            "VT":"Vermont",
                            "VA":"Virginia",
                            "WA":"Washington",
                            "WV":"West Virginia",
                            "WI":"Wisconsin",
                            "WY":"Wyoming",
                            "DC": "District of Columbia",
                            "AS": "American Samoa",
                            "GU": "Guam",
                            "MP": "Northern MAriana Isdlands",
                            "PR": "Puerto Rico",
                            "United States Minor Outlying Islands": "UM",
                            "VI": "U.S. Virgin Islands",
                            "RMI": "The Marshall Islands",
                            "FSM": "FEDERATED STATES OF MICRONESIA RELATIONS",
                            "PW": "Palau"
                           }, inplace = True)

#dropping everything that is not a part of the 50 major states
df3 = df3[df3["state"].str.contains("American Samoa") == False]
df3 = df3[df3["state"].str.contains("Puerto Rico") == False]
df3 = df3[df3["state"].str.contains("FEDERATED STATES OF MICRONESIA RELATIONS
df3 = df3[df3["state"].str.contains("Guam") == False]
df3 = df3[df3["state"].str.contains("Northern MAriana Isdlands") == False]
df3 = df3[df3["state"].str.contains("Palau") == False]
df3 = df3[df3["state"].str.contains("The Marshall Islands") == False]
df3 = df3[df3["state"].str.contains("U.S. Virgin Islands") == False]

#adjustdates to datetime64 type
df3['submission_date'] = pd.to_datetime(df3['submission_date'])

df3
df3.info()
```

Out[12]:

	submission_date	state	tot_cases	new_case	tot_death	new_death
0	2022-01-28	Georgia	2346518	18785	32868	150
2	2022-01-28	Florida	5501599	22705	64647	7
3	2022-01-28	Utah	875251	6166	4107	10
4	2022-01-28	Hawaii	208253	1848	1152	5
6	2022-01-28	Oklahoma	963655	10539	12044	0
...	...	...	...	...	...	...
44274	2020-01-22	Florida	0	0	0	0
44275	2020-01-22	Ohio	0	0	0	0
44277	2020-01-22	New Mexico	0	0	0	0
44278	2020-01-22	Iowa	0	0	0	0
44279	2020-01-22	Pennsylvania	0	0	0	0

38376 rows × 6 columns

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 38376 entries, 0 to 44279
Data columns (total 6 columns):
 #   Column           Non-Null Count  Dtype         
---  ------           --------------  -----         
 0   submission_date  38376 non-null  datetime64[ns]
 1   state            38376 non-null  object        
 2   tot_cases        38376 non-null  int64         
 3   new_case         38376 non-null  int64         
 4   tot_death        38376 non-null  int64         
 5   new_death        38376 non-null  int64         
dtypes: datetime64[ns](1), int64(4), object(1)
memory usage: 2.0+ MB
```

df4 was used to plot the totals such as total cases and total deaths by state

In [14]:

```
#create df of final date, to gather total numbers by end -->
#true tot_cases/tot_death on final date
df4 = df3[df3['submission_date'] == '2022-01-28']
df4.head()
```

Out[14]:

	submission_date	state	tot_cases	new_case	tot_death	new_death
0	2022-01-28	Georgia	2346518	18785	32868	150
2	2022-01-28	Florida	5501599	22705	64647	7
3	2022-01-28	Utah	875251	6166	4107	10
4	2022-01-28	Hawaii	208253	1848	1152	5
6	2022-01-28	Oklahoma	963655	10539	12044	0

newdf was used to groupby state in order to plot totals without any duplicates of states, such as New York.

In [15]:

```
newdf = df4.groupby('state').sum().reset_index()
newdf.head()
```

Out[15]:

	state	tot_cases	new_case	tot_death	new_death
0	Alabama	1206308	10748	17086	39
1	Alaska	205241	5689	1052	4
2	Arizona	1829406	15610	26001	69
3	Arkansas	768061	5660	9616	20
4	California	8213786	76729	78825	254

df5 wascreated to add the percentage of deaths per cases column to the newdf

In [16]:

```
#Calculate percentage of deaths from total cases and total deaths ((total dea

df5 = newdf.copy()
df5["percentage_of_deaths_by_cases"] = (df5['tot_death'] / df5['tot_cases'] *
df5.head()
```

Out[16]:

	state	tot_cases	new_case	tot_death	new_death	percentage_of_deaths_by_cases
0	Alabama	1206308	10748	17086	39	1.416
1	Alaska	205241	5689	1052	4	0.513
2	Arizona	1829406	15610	26001	69	1.421
3	Arkansas	768061	5660	9616	20	1.252
4	California	8213786	76729	78825	254	0.960

dfgrouped was used to group by state and submission date in order to plot, state specifically, new cases and new deaths by date

In [17]:
```
#grouping df3 by submission date and state
dfgrouped = df3.groupby(['state', 'submission_date']).sum().reset_index()
dfgrouped.head()
```

Out[17]:

	state	submission_date	tot_cases	new_case	tot_death	new_death
0	Alabama	2020-01-22	0	0	0	0
1	Alabama	2020-01-23	0	0	0	0
2	Alabama	2020-01-24	0	0	0	0
3	Alabama	2020-01-25	0	0	0	0
4	Alabama	2020-01-26	0	0	0	0

states_few was created to group the 10 states that I felt told the best story of my data

In [19]:
```
#create few states so can create scatter of few states
#Top 5 states by total cases and top 5 states by highest ratio of deaths
states = ['California','New York','Florida','Pennsylvania','Texas',
          'Mississippi', 'New Jersey', 'Michigan', 'Connecticut', 'Arizona']
states_few = dfgrouped[dfgrouped['state'].isin(states)]
states_few
```

Out[19]:

	state	submission_date	tot_cases	new_case	tot_death	new_death
1476	Arizona	2020-01-22	0	0	0	0
1477	Arizona	2020-01-23	0	0	0	0
1478	Arizona	2020-01-24	0	0	0	0
1479	Arizona	2020-01-25	0	0	0	0
1480	Arizona	2020-01-26	1	1	0	0
...	...	...	...	...	...	...
32467	Texas	2022-01-24	5973164	38924	76904	29
32468	Texas	2022-01-25	6018220	45056	77058	154
32469	Texas	2022-01-26	6048954	30734	77321	263
32470	Texas	2022-01-27	6083750	34796	77555	234
32471	Texas	2022-01-28	6122432	38682	77780	225

7380 rows × 6 columns

df6 was used to group the dataset y date in order to create Date vs new cases and new deaths plots

In [21]:

```
#group by submission dates, in order to show cases per day

df6 = df3.groupby(by = 'submission_date').aggregate(np.sum)
df6.index.name = 'Date'
df6 = df6.reset_index()
df6
```

Out[21]:

	Date	tot_cases	new_case	tot_death	new_death
0	2020-01-22	0	0	0	0
1	2020-01-23	1	1	0	0
2	2020-01-24	2	1	0	0
3	2020-01-25	2	0	0	0
4	2020-01-26	3	1	0	0
...	...	...	...	...	...
733	2022-01-24	71391336	1077704	863638	2482
734	2022-01-25	71877826	486490	866751	3113
735	2022-01-26	72439170	561344	869829	2810
736	2022-01-27	73010190	571020	872453	2624
737	2022-01-28	73531094	520904	875755	3181

738 rows × 5 columns

df7 adds a column to df6: percentage_of_new_deaths_by_new_cases

In [22]:

```
#ratio of new deaths/new cases * 100

df7 = df6.copy()
df7["percentage_of_new_deaths_by_new_cases"] = (df6['new_death'] / df6['new_c
df7
```

Out[22]:

	Date	tot_cases	new_case	tot_death	new_death	percentage_of_new_deaths_by_new_c
0	2020-01-22	0	0	0	0	
1	2020-01-23	1	1	0	0	0
2	2020-01-24	2	1	0	0	0
3	2020-01-25	2	0	0	0	
4	2020-01-26	3	1	0	0	0
...	...	...	...	...	...	
733	2022-01-24	71391336	1077704	863638	2482	0
734	2022-01-25	71877826	486490	866751	3113	0
735	2022-01-26	72439170	561344	869829	2810	0
736	2022-01-27	73010190	571020	872453	2624	0
737	2022-01-28	73531094	520904	875755	3181	

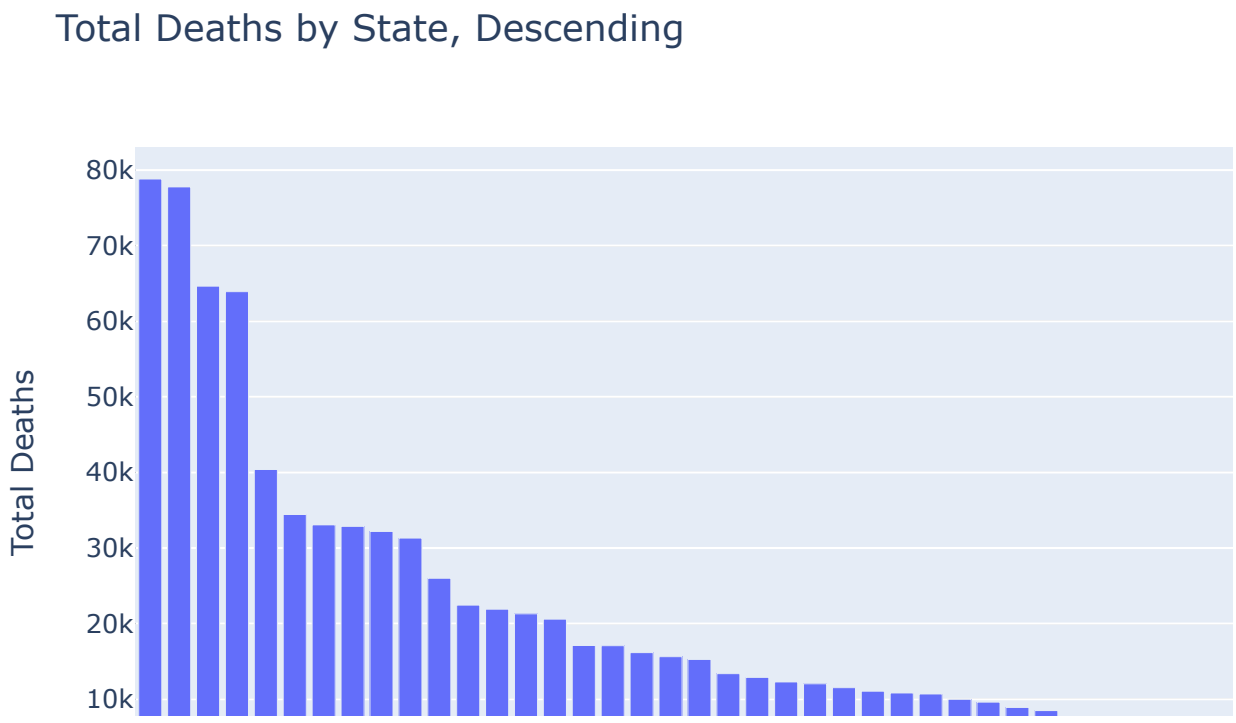
738 rows × 6 columns

Codes that showed most important findings

The first important finding of my EDA was the plot of total deaths by State, in descending order. This was able to show the total number of deaths each state had throughout the pandemic and put it in perspective to see what states lost the most lives.

In [24]:

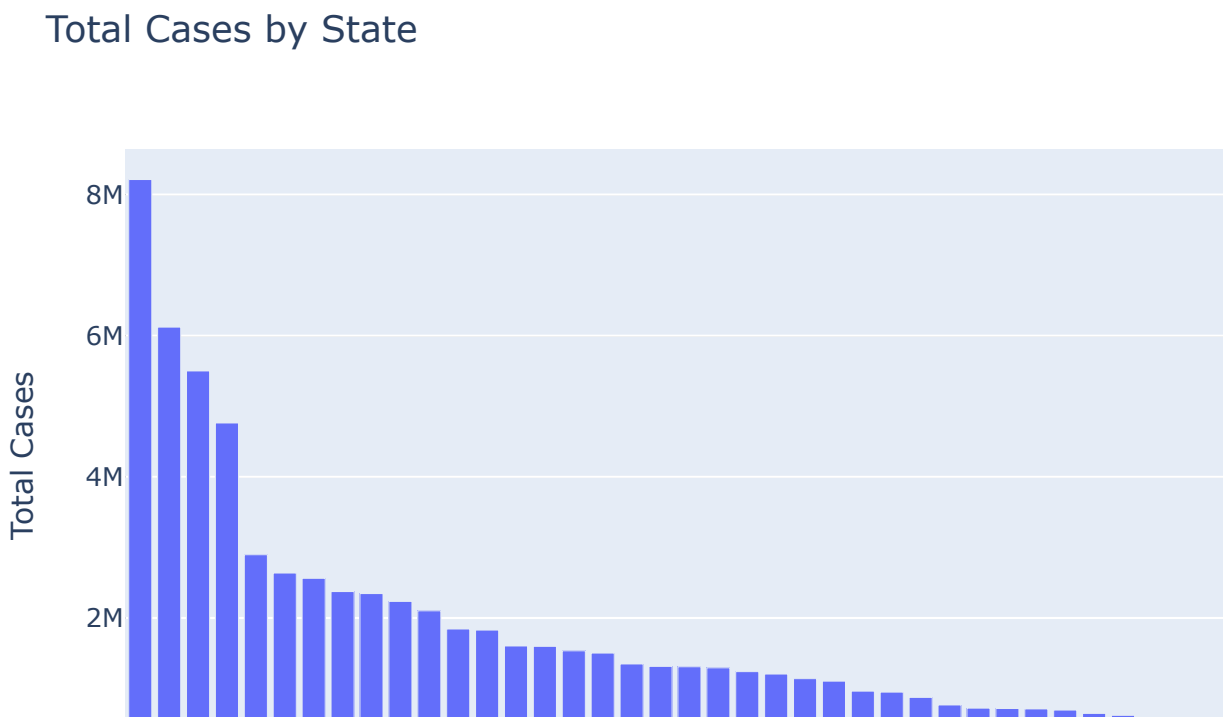
```
#descending states by total death, to show which state had the most, and leas
#cumulative deaths by January 28, 2022.

tot_death_desc = newdf.sort_values(by = 'tot_death', ascending = False)
bar_desc_tot_death = px.bar(tot_death_desc, x = 'state', y = 'tot_death',titl
                            labels = dict(tot_death = "Total Deaths",
                            state = "State"))
bar_desc_tot_death.show()
```



Placing these in descending order was important as we are now able to see that California, Texas, Florida, New York, and Pennsylvania had the most deaths due to covid-19 over the past two years. While Vermont, Alaska, and Hawaii, had the least.

The next significant plot was Total cases by State in descending order

In [25]:
```
# Most cumulative cases by state, as of January 28, 2022
# total cases by state, descending barplot

tot_cases_desc = newdf.sort_values(by = 'tot_cases', ascending = False)
bar_tot_cases_desc = px.bar(tot_cases_desc, x = 'state', y = 'tot_cases',titl
                            labels = dict(tot_cases = "Total Cases",
                            state = "State"))
bar_tot_cases_desc.show()
```



This is significant as it shows how many cases each state had over the pandemic. We can see that California, Texas, Florida, New York, and Pennsylvania are all at the top of the plot, with the most number of cases. This lead me to beleive that there is a correlation between total cases and total deaths.

checking for correlation between total cases and total deaths

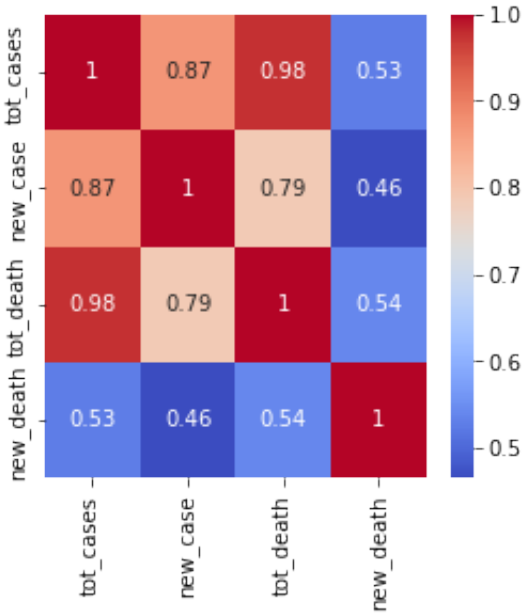
In [28]:
```
#checking correlations
newdf.corr().style.background_gradient(cmap='coolwarm')

#create correlation heat map of total deaths and total cases per state

plt.figure(figsize=(4,4))
seaborn.heatmap(newdf.corr(), annot = True, cmap = 'coolwarm')
```

Out[28]:

	tot_cases	new_case	tot_death	new_death
tot_cases	1.000000	0.871048	0.976014	0.526586
new_case	0.871048	1.000000	0.787194	0.464684
tot_death	0.976014	0.787194	1.000000	0.538976
new_death	0.526586	0.464684	0.538976	1.000000

Out[28]:
```
<Figure size 288x288 with 0 Axes>
```

Out[28]:
```
<AxesSubplot:>
```



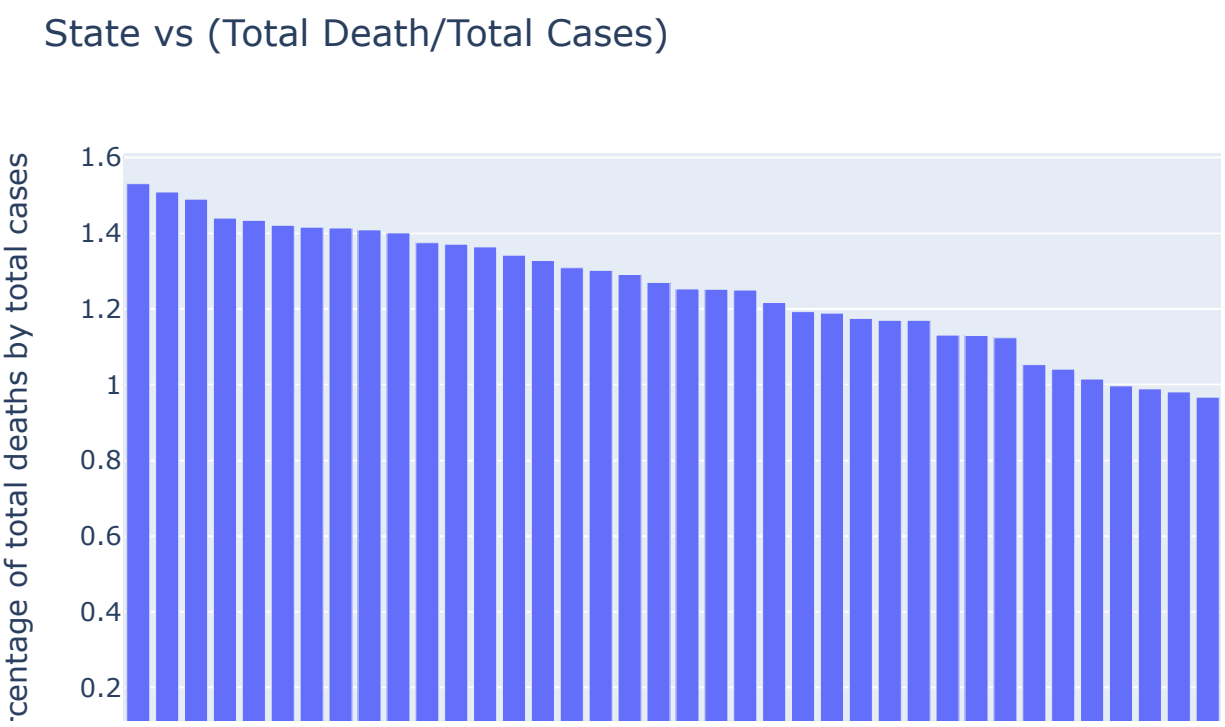
Heat map and correlation table show strong correlation between total deaths and total cases.

Although there is a correlation between total deaths and total cases, we cannot say there is a direct causation. As the population of California is still much higher than states like Vermont. A better way to get a deeper understanding of the overall picture is to get the percentage of death rates of total cases per each state to see what states really did end up "suffering" the most

Since checking the total deaths and cases cannot show a direct correlation with which states suffered the most, due to population size, it was best to look at the ratio of total deaths by total cases to see which states did suffer the most in terms of percentages

In [30]:

```
#plot in descending order to see which state suffered more in relation
#to their number of cases.

relation_death_to_case_per_state = df5.sort_values(by = 'percentage_of_deaths
                                                   ascending = False)
bar_death_per_case_desc = px.bar(relation_death_to_case_per_state,
                                 x = 'state', y = 'percentage_of_deaths_by_ca
                                 title = "State vs (Total Death/Total Cases)"
                            labels = dict(percentage_of_deaths_by_cases = "Pe
                            state = "State"))
bar_death_per_case_desc.show()

#descending order table; by percentage of deaths

relation_death_to_case_per_state
```



Out[30]:

	state	tot_cases	new_case	tot_death	new_death	percentage_of_deaths_by_case
38	Pennsylvania	2637717	15583	40394	137	1.53
24	Mississippi	717666	5533	10831	25	1.50
30	New Jersey	2102227	10118	31320	112	1.49
22	Michigan	2235180	14999	32197	25	1.44
6	Connecticut	696070	2684	9985	77	1.43
2	Arizona	1829406	15610	26001	69	1.42
0	Alabama	1206308	10748	17086	39	1.41
18	Louisiana	1105273	6483	15631	61	1.41
20	Maryland	949880	3011	13387	53	1.40
10	Georgia	2346518	18785	32868	150	1.40
28	Nevada	648088	4393	8914	39	1.37
21	Massachusetts	1598451	8149	21909	70	1.37
31	New Mexico	470513	5269	6417	26	1.36
32	New York	4764000	8558	63921	85	1.34
14	Indiana	1604072	17067	21301	136	1.32
48	West Virginia	438889	4668	5743	46	1.30
25	Missouri	1314435	9106	17111	9	1.30
35	Ohio	2562412	9440	33071	582	1.29
43	Texas	6122432	38682	77780	225	1.27
26	Montana	238801	2834	2993	3	1.25
3	Arkansas	768061	5660	9616	20	1.25
36	Oklahoma	963655	10539	12044	0	1.25
42	Tennessee	1844780	17692	22452	73	1.21
15	Iowa	712288	8922	8501	0	1.19
13	Illinois	2897174	15453	34439	141	1.18

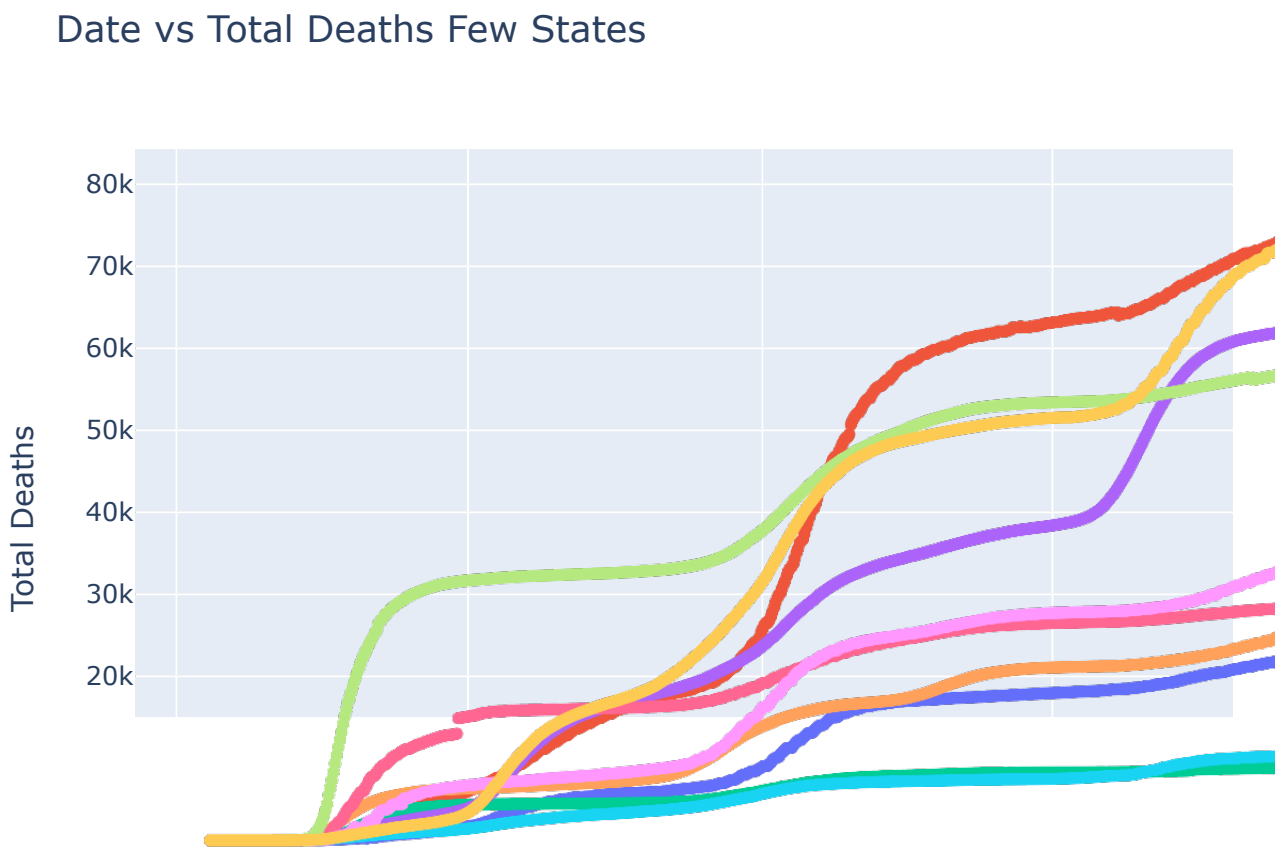
9	Florida	5501599	22705	64647	7	1.17
12	Idaho	376095	3653	4400	35	1.17
41	South Dakota	225383	1145	2637	9	1.17
40	South Carolina	1349276	10892	15266	85	1.13
17	Kentucky	1140887	15706	12890	34	1.13
50	Wyoming	144526	1397	1625	0	1.12
46	Virginia	1535349	9743	16168	41	1.05
16	Kansas	722824	12986	7522	134	1.04
7	Delaware	246037	1307	2498	4	1.01
19	Maine	174217	1266	1737	4	0.99
8	District of Columbia	129817	0	1284	0	0.98
37	Oregon	620652	7431	6086	19	0.98
39	Rhode Island	341407	1836	3302	14	0.96
4	California	8213786	76729	78825	254	0.96
34	North Dakota	221025	2065	2093	3	0.94
5	Colorado	1240361	7083	11061	56	0.89
23	Minnesota	1309665	14548	11532	43	0.88
33	North Carolina	2374866	22631	20595	78	0.86
27	Nebraska	435358	0	3666	0	0.84
47	Washington	1294498	13483	10699	52	0.82
49	Wisconsin	1503420	8180	12291	75	0.81
29	New Hampshire	272492	2429	2205	12	0.80
11	Hawaii	208253	1848	1152	5	0.55
45	Vermont	94513	0	503	0	0.53
1	Alaska	205241	5689	1052	4	0.51
44	Utah	875251	6166	4107	10	0.46

We can now clearly see that Pennsylvania, suffered the most with 1.531% of its total cases resulting in deaths. We also know from prior plots and tables that Pennsylvania also was amongst the top 5 total cases and total deaths. Furthermore, states like Mississippi, New Jersey, and Michigan, although not in highest cases and deaths, still suffered the most as they were amongst the top 5 in ratio of deaths per cases.

I then chose the states with the top 5 deaths and top 5 ratios to plot on a time line next to each other to see their trends throughout the pandemic

In [31]:

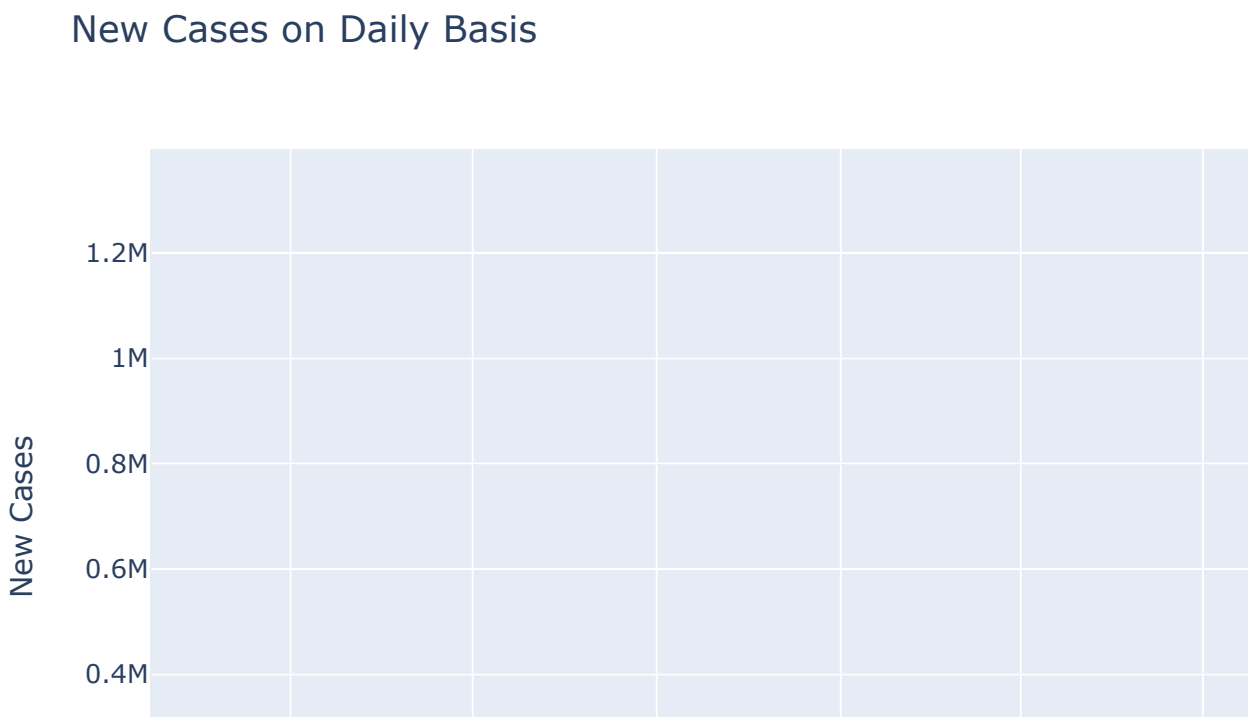
```
#show in a few states at once, dates vs total deaths scatter:
#this is a scatter plot of date by total death of the top 5 states by cases a

scatter_few_states = px.scatter(states_few,
            x = 'submission_date', y = 'tot_death', color = 'state',
              title = "Date vs Total Deaths Few States",
                            labels = dict(submission_date = "Date",
                            tot_death = "Total Deaths"))
scatter_few_states.show()
```



This was significant because it shows us the trends of the states with the most deaths and highest suffering throughout the pandemic. As well as how the increased over time. We can see that for California, Texas, Florida, and New York, things really started to increase tremendously starting in the end of 2021. While all states saw a jump in the start of 2021.

To more accurately show the trend of new cases on a daily basis I plotted the cumulative new cases on a daily basis for all of the states on one plot

In [32]:
```
#new cases on daily basis

new_cases_daily = px.area(df6, x = 'Date', y = 'new_case',
              title = "New Cases on Daily Basis",
             labels = dict(new_case = "New Cases"))
new_cases_daily.show()
```



This was important as we can now see the overall story of this pandemic within the United States over the past two years. We can identify that hill over Jan 2021 was the Beta variant, while the hill inbetween Jul 2021 and Oct 2021 is the Delta variant, and the large peaknear Jan 2022 is the Omicron variant. We can also have a hunch that Omicron was the most contageous. We cannot entirely conclude that from this data and graph alone. We would need to know What variant exactly each person tested positive for, to do that.

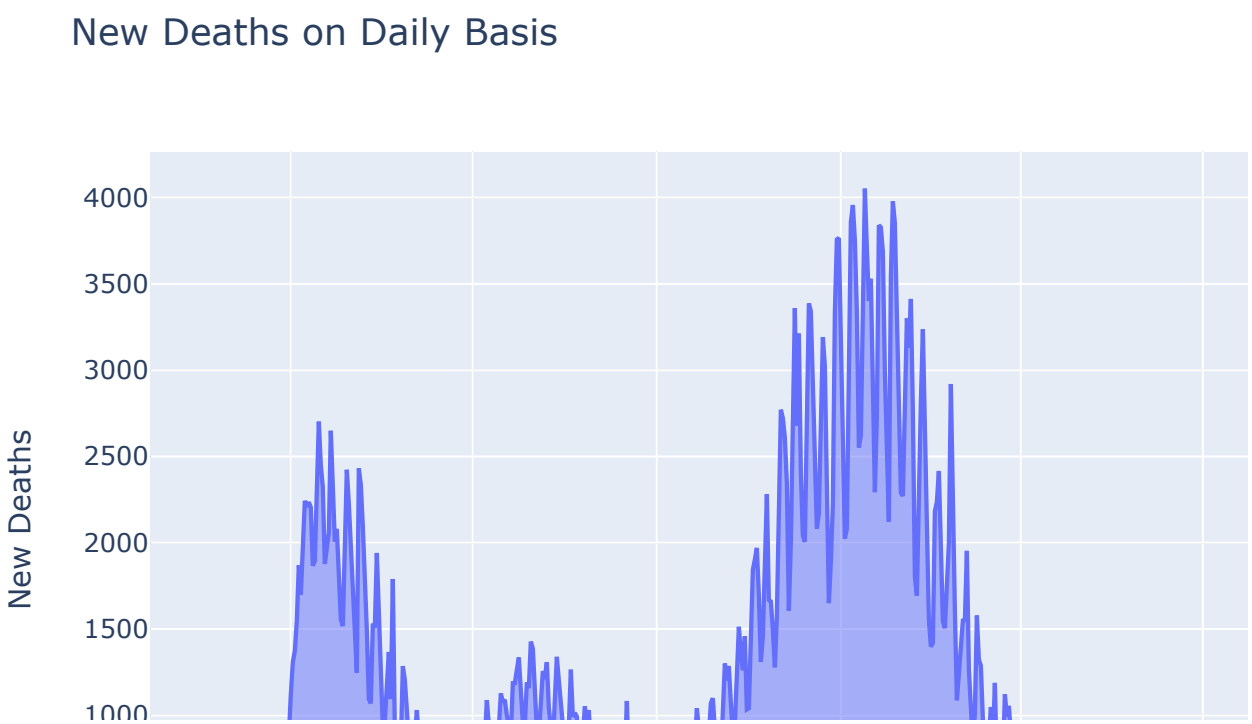
further information on when each variant was introduced to the United States can be found here:

https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/

Although it is not possible to conclude which variant was the most deadly, I felt like it would still be useful to know how many deaths occured on a dasily basis over the timeline of the pandemic. As it might be able to lead us to new hunches on how the people of the USA reacted to each variant.

In [34]:

```
#New deaths daily to show severity of different varients

new_deaths_daily = px.area(df6, x = 'Date', y = 'new_death',
              title = "New Deaths on Daily Basis",
             labels = dict(new_death = "New Deaths"))
new_deaths_daily.show()
```

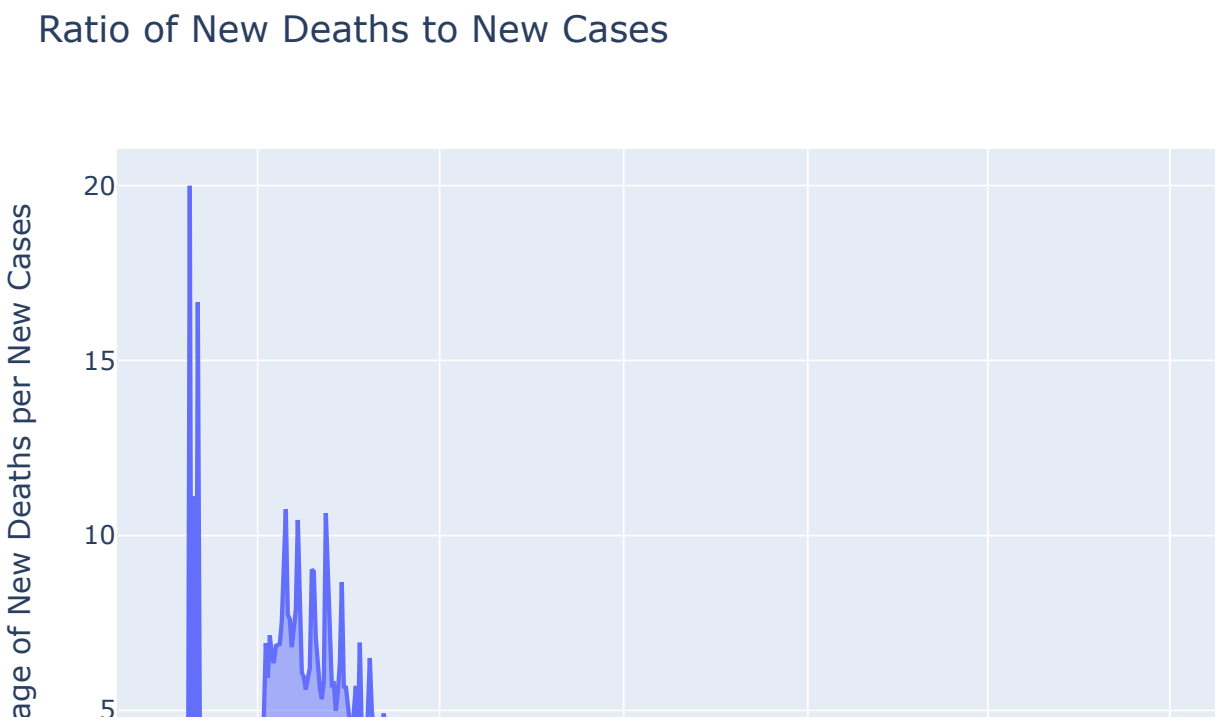


Here we can see that it looks like the delta variant caused the most deaths. Yet we cannot say it is the most deadly, as the Beta and Alpha variants at the beginning were not far off, and we did not have a vaccine back then, while during the Delta variant, a good number of high risk people were vaccinated by then. In addition to this, based on this data alone, it would be hard to determine which variant is the most deadly as we would need more information. Furthermore, the Omicron variant had significantly more cases, yet we did not have the most deaths during that time. To get a better look as to what variants were the most lethal, we could take a etter look at the percentage of new deaths per new cases.

Looking at the ratio of new deaths per new cases to give a more clear answer on what variant was th emost deadly.

In [35]:

```
#graph of ratio of new deaths per new cases

new_deaths_per_cases_daily = px.area(df7, x = 'Date', y = 'percentage_of_new_
              title = "Ratio of New Deaths to New Cases",
             labels = dict(percentage_of_new_deaths_by_new_cases = "Percentag
new_deaths_per_cases_daily.show()
```



Looking at this plot we can see that the Alpha and Beta variants actually killed the most people in ratio to the number of cases we were having within the USA. Although it is compelling to say that these variants were the most deadly because they caused the most deaths, in proportion, we cannot conclude that from this data alone. We can just use this as a hunch or theory. As these results could be due to a multitude of variables. Such as limited testing early on during the pandemic. Furthermore, we can see that the ratio of deaths during the Omicron variant period was very low, even though there was an abundance of testing as well as more people vaccinated.

Summary:

The purpose of this project was to show a timeline of how each state was effected through the pandemic. This was done through visualizing and analyzing the total deaths, total cases, new deaths, new cases, and dates on a cumulative and independent basis for the main fifty

states of the USA.

My hypothesis was that states with looser mask mandates, such as Florida and Texas were going to be amongst the states that suffered the most. Initially when looking a the total number of cases and deaths, this seemed to be the case. However, when looking at the raito of deaths by cases for each state, we can see that Florida and Texas are not amongst the top five states that suffered through the pandemic. Shockingly, it seemed to have been Pennsylvania, as it was amongst the top five in total cases, total deaths, and number one when it came to the ratio of deaths per cases. At the same time, we saw that California, even though it had a high number of cases and deaths, was at the bottom quarter of states that suffered the most. These contradicting results could be due to a number of different factors. It could have been that the stricter mask mandate in California was helping, yet, Pennsylvania, which also had a relatively strict mask mandate was at the top of the list when it came to suffering. I beleive that there is one key red flag that stood out to me the most upoon completing this project. The fact that this data sent to the CDC was voluntary. This meant that each state could have biases to skew their data as they wished before submitting the data to the CDC. It also meant they could report as they want and as much as they want.

My second hypothesis/hunch was that this data could visually represent the story of the pandemic within the USA. My hunch was that looking at the data, one would be able to tell when each variant was entering the United States, and that we would be able to see how we reacted to the variants as a whole. This is evident through the data, we can see when we first started experiencing the alpha and delta varients in the end of 2020, with the delta variant eing right near the end of 2021, and ust before the Omicron variant. We can even see the surge of cases at the beginning of the pandamic.

Overall I was able to take the data and visualize the trends of covid throughout the pandamic, along with the varaints and their spikes that came with them. Furthermore, based on the CDC's volunatry data, the ratio of total deaths per total cases disproved my theory that states with loser mask mandates would have suffered the most. As the plots indicate, Pennsylvania is actually the state that suffered the most. Furthermore, although it is difficult to make a decisive conclusion, the ratio of deaths per cases helped theorize that the Delta variant could have been the most deadly covid19 variant within the USA to this date.

Possible next steps would be to compare the CDC's data to other instiutions, such as John Hopkins University, and a few other instiutions, to see how accurate the voluntary data. Furthermore, we could take this one step forward and obtain what variant each individual tested positive for (if possible).

In [ ]: