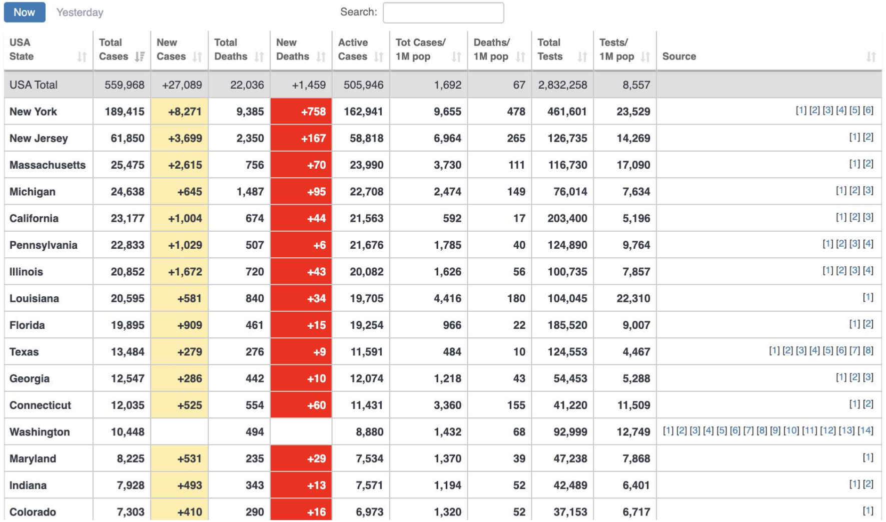
Working with CSV Files in Python with Jupyter Notebook

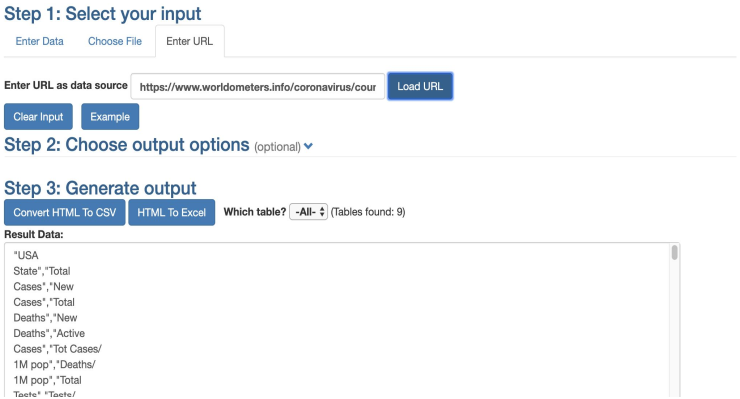
Yvette Green
CS 5001
Spring 2020

COVID-19 Data

- The COVID-19 data was obtained from Worldometers website, self-described as a website “run by an international team of developers, researchers, and volunteers with the goal of making world statistics available.”
- Because the data was in HTML format, I used CSV conversion website to convert the table from HTML to a CSV file.



USA State	Total Cases	New Cases	Total Deaths	New Deaths	Active Cases	Tot Cases/ 1M pop	Deaths/ 1M pop	Total Tests	Tests/ 1M pop	Source
USA Total	559,968	+27,089	22,036	+1,459	505,946	1,692	67	2,832,258	8,557	
New York	189,415	+8,271	9,385	+758	162,941	9,655	478	461,601	23,529	[1] [2] [3] [4] [5] [6]
New Jersey	61,850	+3,699	2,350	+167	58,818	6,964	265	126,735	14,269	[1] [2]
Massachusetts	25,475	+2,615	756	+70	23,990	3,730	111	116,730	17,090	[1] [2]
Michigan	24,638	+645	1,487	+95	22,708	2,474	149	76,014	7,634	[1] [2] [3]
California	23,177	+1,004	674	+44	21,563	592	17	203,400	5,196	[1] [2] [3]
Pennsylvania	22,833	+1,029	507	+6	21,676	1,785	40	124,890	9,764	[1] [2] [3] [4]
Illinois	20,852	+1,672	720	+43	20,082	1,626	56	100,735	7,857	[1] [2] [3] [4]
Louisiana	20,595	+581	840	+34	19,705	4,416	180	104,045	22,310	[1]
Florida	19,895	+909	461	+15	19,254	966	22	185,520	9,007	[1] [2]
Texas	13,484	+279	276	+9	11,591	484	10	124,553	4,467	[1] [2] [3] [4] [5] [6] [7] [8]
Georgia	12,547	+286	442	+10	12,074	1,218	43	54,453	5,288	[1] [2] [3]
Connecticut	12,035	+525	554	+60	11,431	3,360	155	41,220	11,509	[1] [2]
Washington	10,448		494		8,880	1,432	68	92,999	12,749	[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14]
Maryland	8,225	+531	235	+29	7,534	1,370	39	47,238	7,868	[1]
Indiana	7,928	+493	343	+13	7,571	1,194	52	42,489	6,401	[1] [2]
Colorado	7,303	+410	290	+16	6,973	1,320	52	37,153	6,717	[1]



Step 1: Select your input

Enter URL as data source: https://www.worldometers.info/coronavirus/cour

Step 2: Choose output options (optional)

Step 3: Generate output

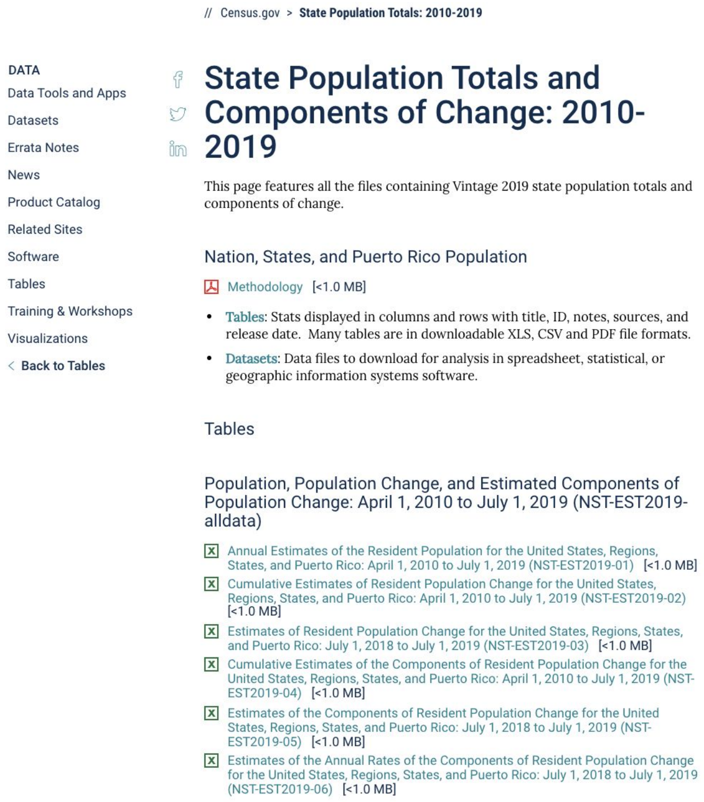
Which table? -All- (Tables found: 9)

Result Data:

```
"USA
State","Total
Cases","New
Cases","Total
Deaths","New
Deaths","Active
Cases","Tot Cases/
1M pop","Deaths/
1M pop","Total
Tests","Tests/
```

US Census Population Data

- The US population data was obtained from the United States Census Bureau website (census.gov), which has downloadable tables and datasets of population totals and population change estimates.
- For this project, I used the table with 2019 population estimates for each U.S. state.



State Population Totals and Components of Change: 2010-2019

This page features all the files containing Vintage 2019 state population totals and components of change.

Nation, States, and Puerto Rico Population

Methodology [<1.0 MB]

- Tables: Stats displayed in columns and rows with title, ID, notes, sources, and release date. Many tables are in downloadable XLS, CSV and PDF file formats.
- Datasets: Data files to download for analysis in spreadsheet, statistical, or geographic information systems software.

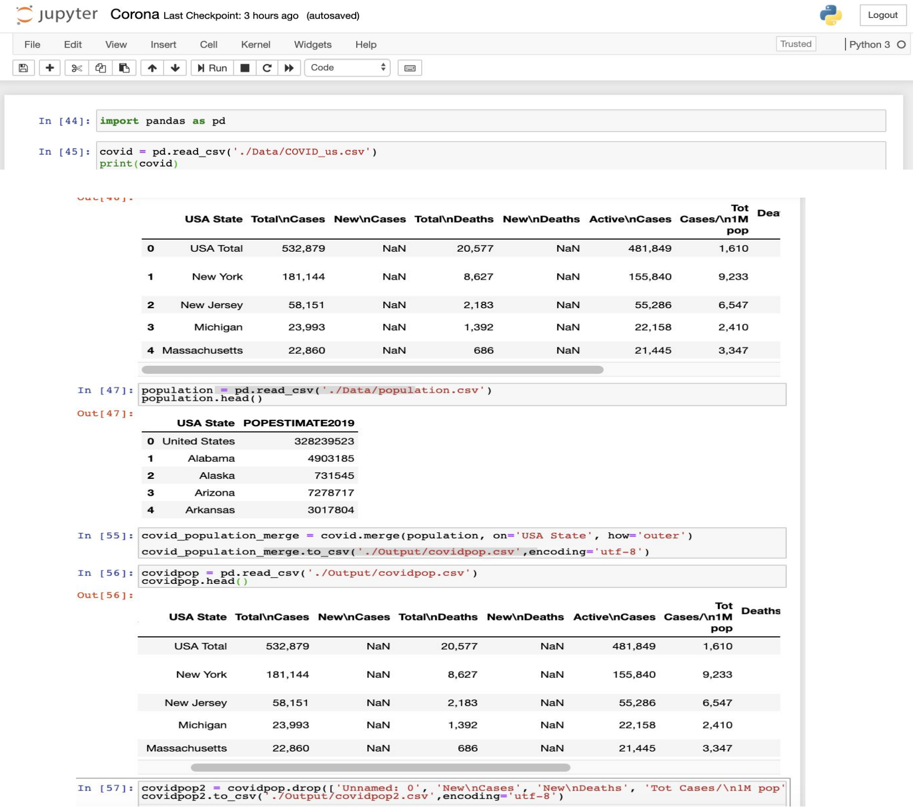
Tables

Population, Population Change, and Estimated Components of Population Change: April 1, 2010 to July 1, 2019 (NST-EST2019-alldata)

- Annual Estimates of the Resident Population for the United States, Regions, States, and Puerto Rico: April 1, 2010 to July 1, 2019 (NST-EST2019-01) [<1.0 MB]
- Cumulative Estimates of Resident Population Change for the United States, Regions, States, and Puerto Rico: April 1, 2010 to July 1, 2019 (NST-EST2019-02) [<1.0 MB]
- Estimates of Resident Population Change for the United States, Regions, States, and Puerto Rico: July 1, 2018 to July 1, 2019 (NST-EST2019-03) [<1.0 MB]
- Cumulative Estimates of the Components of Resident Population Change for the United States, Regions, States, and Puerto Rico: April 1, 2010 to July 1, 2019 (NST-EST2019-04) [<1.0 MB]
- Estimates of the Components of Resident Population Change for the United States, Regions, States, and Puerto Rico: July 1, 2018 to July 1, 2019 (NST-EST2019-05) [<1.0 MB]
- Estimates of the Annual Rates of the Components of Resident Population Change for the United States, Regions, States, and Puerto Rico: July 1, 2018 to July 1, 2019 (NST-EST2019-06) [<1.0 MB]

Merging CSV files in Python

- I saved the CSV files in a Data folder.
- Using Jupyter Notebook, I merged the files and removed some of the columns.



```
In [44]: import pandas as pd

In [45]: covid = pd.read_csv('./Data/COVID_us.csv')
         print(covid)
```

	USA State	Total\nCases	New\nCases	Total\nDeaths	New\nDeaths	Active\nCases	Tot Cases/\n1M pop	Dea
0	USA Total	532,879	NaN	20,577	NaN	481,849	1,610	
1	New York	181,144	NaN	8,627	NaN	155,840	9,233	
2	New Jersey	58,151	NaN	2,183	NaN	55,286	6,547	
3	Michigan	23,993	NaN	1,392	NaN	22,158	2,410	
4	Massachusetts	22,860	NaN	686	NaN	21,445	3,347	

```
In [47]: population = pd.read_csv('./Data/population.csv')
         population.head()
Out[47]:
```

	USA State	POPESTIMATE2019
0	United States	328239523
1	Alabama	4903185
2	Alaska	731545
3	Arizona	7278717
4	Arkansas	3017804

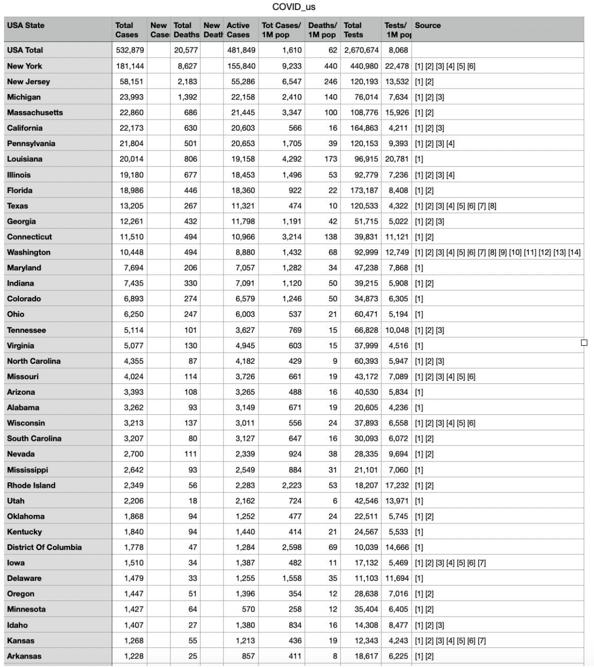
```
In [55]: covid_population_merge = covid.merge(population, on='USA State', how='outer')
         covid_population_merge.to_csv('./Output/covidpop.csv',encoding='utf-8')

In [56]: covidpop = pd.read_csv('./Output/covidpop.csv')
         covidpop.head()
Out[56]:
```

USA State	Total\nCases	New\nCases	Total\nDeaths	New\nDeaths	Active\nCases	Tot Cases/\n1M pop	Deaths
USA Total	532,879	NaN	20,577	NaN	481,849	1,610	
New York	181,144	NaN	8,627	NaN	155,840	9,233	
New Jersey	58,151	NaN	2,183	NaN	55,286	6,547	
Michigan	23,993	NaN	1,392	NaN	22,158	2,410	
Massachusetts	22,860	NaN	686	NaN	21,445	3,347	

```
In [57]: covidpop2 = covidpop.drop(['Unnamed: 0', 'New\nCases', 'New\nDeaths', 'Tot Cases/\n1M pop'
         covidpop2.to_csv('./Output/covidpop2.csv',encoding='utf-8')
```

Original CSV Files

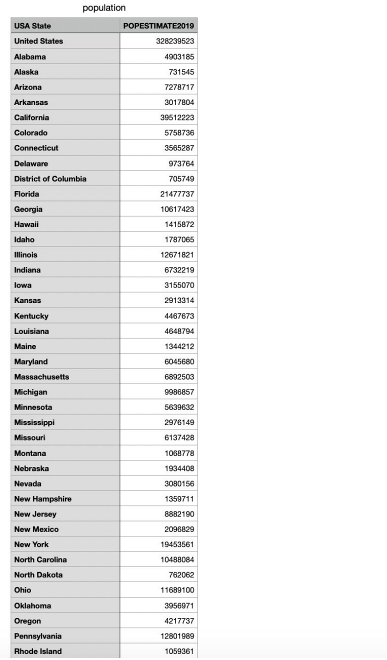
COVID-19 data



COVID_us

USA State	Total Cases	New Case	Total Deaths	New Death	Active Cases	Tot Cases/ 1M pop	Deaths/ 1M pop	Total Tests	Tests/ 1M pop	Source
USA Total	532,879		20,577		481,849	1,610	62	2,670,674	8,068	
New York	181,144		8,627		155,840	9,233	440	440,980	22,478	[1] [2] [3] [4] [5] [6]
New Jersey	58,151		2,183		55,286	6,547	246	120,193	13,532	[1] [2]
Michigan	23,993		1,392		22,158	2,410	140	76,014	7,634	[1] [2] [3]
Massachusetts	22,860		686		21,445	3,347	100	108,776	15,926	[1] [2]
California	22,173		630		20,603	566	16	164,863	4,211	[1] [2] [3]
Pennsylvania	21,804		501		20,653	1,705	39	120,153	9,393	[1] [2] [3] [4]
Louisiana	20,014		806		19,158	4,292	173	96,915	20,781	[1]
Illinois	19,180		677		18,453	1,496	53	92,779	7,236	[1] [2] [3] [4]
Florida	18,986		446		18,360	922	22	173,187	8,408	[1] [2]
Texas	13,205		267		11,321	474	10	120,533	4,322	[1] [2] [3] [4] [5] [6] [7] [8]
Georgia	12,261		432		11,798	1,191	42	51,715	5,022	[1] [2] [3]
Connecticut	11,510		494		10,966	3,214	138	39,831	11,121	[1] [2]
Washington	10,448		494		8,880	1,432	68	92,999	12,749	[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14]
Maryland	7,694		206		7,057	1,282	34	47,238	7,868	[1]
Indiana	7,435		330		7,091	1,120	50	39,215	5,908	[1] [2]
Colorado	6,893		274		6,579	1,246	50	34,873	6,305	[1]
Ohio	6,250		247		6,003	537	21	60,471	5,194	[1]
Tennessee	5,114		101		3,627	769	15	66,828	10,048	[1] [2] [3]
Virginia	5,077		130		4,945	603	15	37,999	4,516	[1]
North Carolina	4,355		87		4,182	429	9	60,393	5,947	[1] [2] [3]
Missouri	4,024		114		3,726	661	19	43,172	7,089	[1] [2] [3] [4] [5] [6]
Arizona	3,393		108		3,265	488	16	40,530	5,834	[1]
Alabama	3,262		93		3,149	671	19	20,605	4,236	[1]
Wisconsin	3,213		137		3,011	556	24	37,893	6,558	[1] [2] [3] [4] [5] [6]
South Carolina	3,207		80		3,127	647	16	30,093	6,072	[1] [2]
Nevada	2,700		111		2,339	924	38	28,335	9,694	[1] [2]
Mississippi	2,642		93		2,549	884	31	21,101	7,060	[1]
Rhode Island	2,349		56		2,283	2,223	53	18,207	17,232	[1] [2]
Utah	2,206		18		2,162	724	6	42,546	13,971	[1]
Oklahoma	1,868		94		1,252	477	24	22,511	5,745	[1] [2]
Kentucky	1,840		94		1,440	414	21	24,567	5,533	[1]
District Of Columbia	1,778		47		1,284	2,598	69	10,039	14,666	[1]
Iowa	1,510		34		1,387	482	11	17,132	5,469	[1] [2] [3] [4] [5] [6] [7]
Delaware	1,479		33		1,255	1,558	35	11,103	11,694	[1]
Oregon	1,447		51		1,396	354	12	28,638	7,016	[1] [2]
Minnesota	1,427		64		570	258	12	35,404	6,405	[1] [2]
Idaho	1,407		27		1,380	834	16	14,308	8,477	[1] [2] [3]
Kansas	1,268		55		1,213	436	19	12,343	4,243	[1] [2] [3] [4] [5] [6] [7]
Arkansas	1,228		25		857	411	8	18,617	6,225	[1] [2]

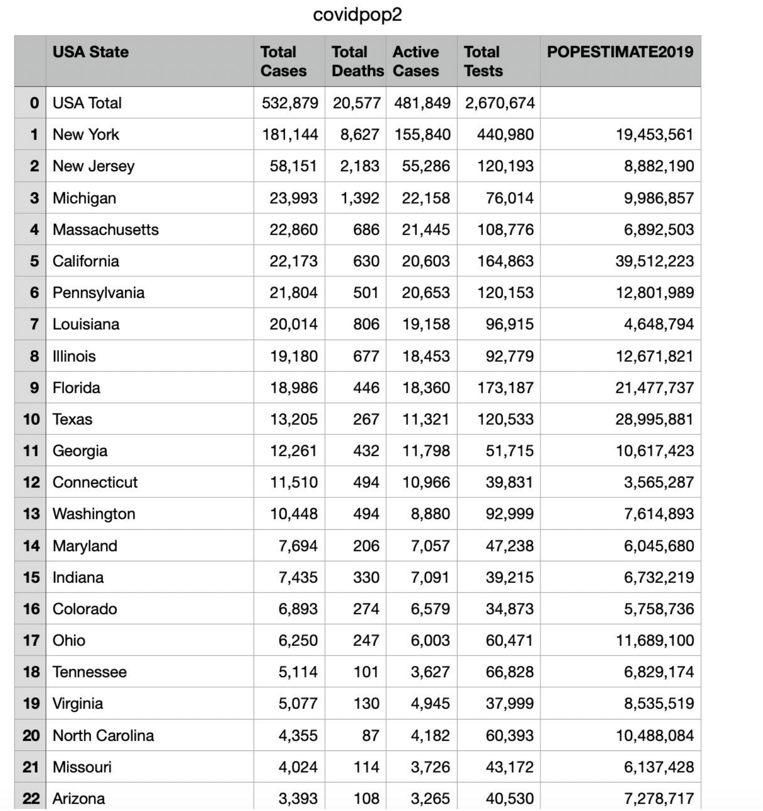
US population data



population

USA State	POPESTIMATE2019
United States	328239523
Alabama	4903185
Alaska	731545
Arizona	7278717
Arkansas	3017804
California	39512223
Colorado	5758736
Connecticut	3565287
Delaware	973764
District of Columbia	705749
Florida	21477737
Georgia	10617423
Hawaii	1415872
Idaho	1787065
Illinois	12671821
Indiana	6732219
Iowa	3155070
Kansas	2913314
Kentucky	4467673
Louisiana	4648794
Maine	1344212
Maryland	6045680
Massachusetts	6892503
Michigan	9986857
Minnesota	5639632
Mississippi	2976149
Missouri	6137428
Montana	1068778
Nebraska	1934408
Nevada	3080156
New Hampshire	1359711
New Jersey	8882190
New Mexico	2096829
New York	19453561
North Carolina	10488084
North Dakota	762062
Ohio	11689100
Oklahoma	3956971
Oregon	4217737
Pennsylvania	12801989
Rhode Island	1059361

Merged File

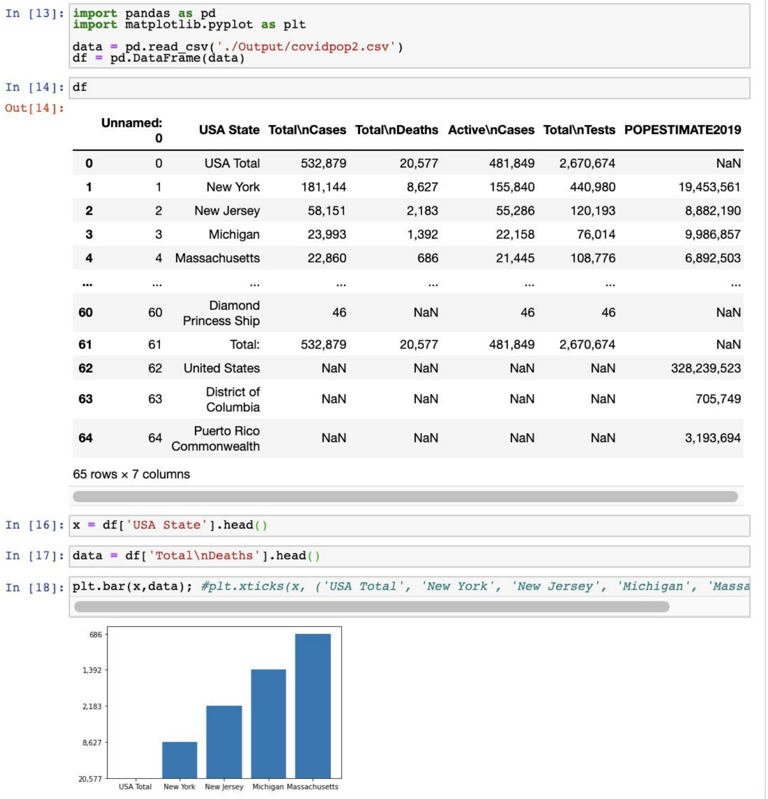


covidpop2

	USA State	Total Cases	Total Deaths	Active Cases	Total Tests	POPESTIMATE2019
0	USA Total	532,879	20,577	481,849	2,670,674	
1	New York	181,144	8,627	155,840	440,980	19,453,561
2	New Jersey	58,151	2,183	55,286	120,193	8,882,190
3	Michigan	23,993	1,392	22,158	76,014	9,986,857
4	Massachusetts	22,860	686	21,445	108,776	6,892,503
5	California	22,173	630	20,603	164,863	39,512,223
6	Pennsylvania	21,804	501	20,653	120,153	12,801,989
7	Louisiana	20,014	806	19,158	96,915	4,648,794
8	Illinois	19,180	677	18,453	92,779	12,671,821
9	Florida	18,986	446	18,360	173,187	21,477,737
10	Texas	13,205	267	11,321	120,533	28,995,881
11	Georgia	12,261	432	11,798	51,715	10,617,423
12	Connecticut	11,510	494	10,966	39,831	3,565,287
13	Washington	10,448	494	8,880	92,999	7,614,893
14	Maryland	7,694	206	7,057	47,238	6,045,680
15	Indiana	7,435	330	7,091	39,215	6,732,219
16	Colorado	6,893	274	6,579	34,873	5,758,736
17	Ohio	6,250	247	6,003	60,471	11,689,100
18	Tennessee	5,114	101	3,627	66,828	6,829,174
19	Virginia	5,077	130	4,945	37,999	8,535,519
20	North Carolina	4,355	87	4,182	60,393	10,488,084
21	Missouri	4,024	114	3,726	43,172	6,137,428
22	Arizona	3,393	108	3,265	40,530	7,278,717

Bar Graph 1 with matplotlib

- To visualize the data, I created a bar graph using matplotlib in Jupyter Notebook.
- I struggled with how to format the numbers on the graph.



```
In [13]: import pandas as pd
         import matplotlib.pyplot as plt

         data = pd.read_csv('./Output/covidpop2.csv')
         df = pd.DataFrame(data)

In [14]: df
Out[14]:
```

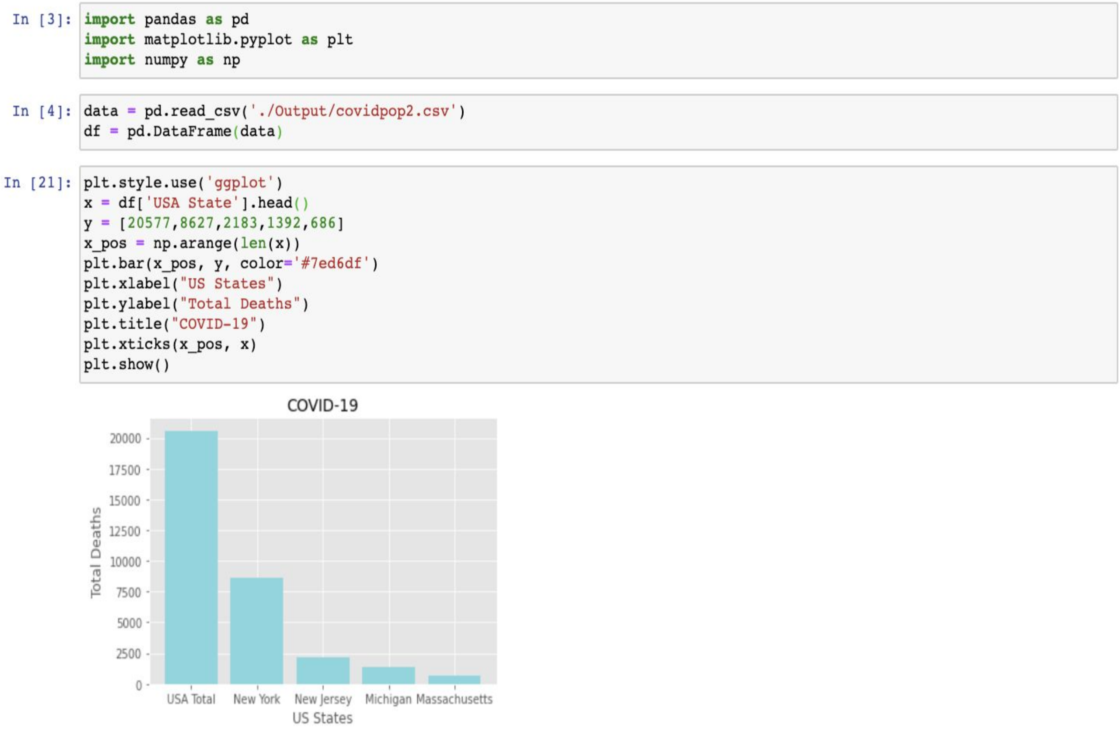
	Unnamed: 0	USA State	Total\nCases	Total\nDeaths	Active\nCases	Total\nTests	POPESTIMATE2019
0	0	USA Total	532,879	20,577	481,849	2,670,674	NaN
1	1	New York	181,144	8,627	155,840	440,980	19,453,561
2	2	New Jersey	58,151	2,183	55,286	120,193	8,882,190
3	3	Michigan	23,993	1,392	22,158	76,014	9,986,857
4	4	Massachusetts	22,860	686	21,445	108,776	6,892,503
...	...	...	...	...	...	...	...
60	60	Diamond Princess Ship	46	NaN	46	46	NaN
61	61	Total:	532,879	20,577	481,849	2,670,674	NaN
62	62	United States	NaN	NaN	NaN	NaN	328,239,523
63	63	District of Columbia	NaN	NaN	NaN	NaN	705,749
64	64	Puerto Rico Commonwealth	NaN	NaN	NaN	NaN	3,193,694

65 rows × 7 columns

```
In [16]: x = df['USA State'].head()

In [17]: data = df['Total\nDeaths'].head()

In [18]: plt.bar(x,data); #plt.xticks(x, ('USA Total', 'New York', 'New Jersey', 'Michigan', 'Massa
```

Bar Graph 2 with matplotlib and NumPy

- I modified the code to properly format the number order while adding labels and different colors.



```
In [3]:  import pandas as pd
         import matplotlib.pyplot as plt
         import numpy as np

In [4]:  data = pd.read_csv('./Output/covidpop2.csv')
         df = pd.DataFrame(data)

In [21]: plt.style.use('ggplot')
         x = df['USA State'].head()
         y = [20577,8627,2183,1392,686]
         x_pos = np.arange(len(x))
         plt.bar(x_pos, y, color='#7ed6df')
         plt.xlabel("US States")
         plt.ylabel("Total Deaths")
         plt.title("COVID-19")
         plt.xticks(x_pos, x)
         plt.show()
```