Covid 19 Vaccine Analysis Milestone 3

October 2, 2024

[9]:
```
import pandas as pd
import requests
import requests
from bs4 import BeautifulSoup
```

[10]:
```
url = 'https://en.wikipedia.org/wiki/List_of_U.S.
 ↪_states_and_territories_by_population'
```

[11]:
```
website_url = requests.get(url).text
soup = BeautifulSoup(website_url,'html.parser')
tables = soup.find_all('table')
table = soup.find("table")
```

[12]:
```
# find all tables
```

[13]:
```
import requests
from bs4 import BeautifulSoup

# URL of the page
url = 'https://en.wikipedia.org/wiki/List_of_U.S.
 ↪_states_and_territories_by_population'
# Send a GET request to the website
response = requests.get(url)

# Parse the HTML content of the page
soup = BeautifulSoup(response.text, 'html.parser')

# Find the table you're interested in, assuming it's the first table
table = soup.find('table')

# Extract data from each row of the table
type(table)
```

[13]:
```
bs4.element.Tag
```

[14]:
```
# find all rows
```

[15]:
```
rows = table.find_all('tr')
```

[16]:
```
type(rows)
```

[16]:
```
bs4.element.ResultSet
```

[17]:
```
# step 1 get all elements from the web and form them into a dataset
```

[18]:
```
popu_data_table = []
for row in rows[2:58]:
    cells = row.find_all('td')
    if len(cells) > 0:
        states = cells[0].get_text().strip()
        population2020 = cells[2].get_text().strip()
        popu_data_table.append([states, population2020])


popu_data = pd.DataFrame(popu_data_table, columns=["States", "Population2020"])
```

[19]:
```
# print dataset
```

[20]:
```
pd.set_option('display.max_rows', None)
```

[21]:
```
print(popu_data)
```

```
                   States Population2020
0              California     39,538,223
1                   Texas     29,145,505
2                 Florida     21,538,187
3                New York     20,201,249
4            Pennsylvania     13,002,700
5                Illinois     12,812,508
6                    Ohio     11,799,448
7                 Georgia     10,711,908
8          North Carolina     10,439,388
9                Michigan     10,077,331
10             New Jersey      9,288,994
11               Virginia      8,631,393
12             Washington      7,705,281
13                Arizona      7,151,502
14              Tennessee      6,910,840
15          Massachusetts      7,029,917
16                Indiana      6,785,528
17               Missouri      6,154,913
18               Maryland      6,177,224
19              Wisconsin      5,893,718
20               Colorado      5,773,714
21              Minnesota      5,706,494
22         South Carolina      5,118,425
23                Alabama      5,024,279
```

```
24                    Louisiana      4,657,757
25                     Kentucky      4,505,836
26                       Oregon      4,237,256
27                     Oklahoma      3,959,353
28                  Connecticut      3,605,944
29                         Utah      3,271,616
30                         Iowa      3,190,369
31                  Puerto Rico      3,285,874
32                       Nevada      3,104,614
33                     Arkansas      3,011,524
34                       Kansas      2,937,880
35                  Mississippi      2,961,279
36                   New Mexico      2,117,522
37                     Nebraska      1,961,504
38                        Idaho      1,839,106
39                West Virginia      1,793,716
40                       Hawaii      1,455,271
41                New Hampshire      1,377,529
42                        Maine      1,362,359
43                      Montana      1,084,225
44                 Rhode Island      1,097,379
45                     Delaware        989,948
46                 South Dakota        886,667
47                 North Dakota        779,094
48                       Alaska        733,391
49         District of Columbia        689,545
50                      Vermont        643,077
51                      Wyoming        576,851
52                    Guam[10]        153,836
53      U.S. Virgin Islands[11]         87,146
54           American Samoa[12]         49,710
55  Northern Mariana Islands[13]         47,329
```

[22]:
```
# step 2 remove numbers and brackets in state column
```

[23]:
```
type(popu_data['States'][0])
```

[23]:
```
str
```

[24]:
```
popu_data['States'] = popu_data['States'].str.replace(r'\d+|\[.*?\]|\(.*?\)',␣
↪'', regex=True)
```

[25]:
```
# step 3 add a column with state abbreviations
```

[26]:
```
state_abbreviations = {
    'Alabama': 'AL', 'Alaska': 'AK', 'Arizona': 'AZ', 'Arkansas': 'AR',␣
↪'California': 'CA',
```

```
    'Colorado': 'CO', 'Connecticut': 'CT', 'Delaware': 'DE', 'Florida': 'FL',␣
↪'Georgia': 'GA',
    'Hawaii': 'HI', 'Idaho': 'ID', 'Illinois': 'IL', 'Indiana': 'IN', 'Iowa':␣
↪'IA',
    'Kansas': 'KS', 'Kentucky': 'KY', 'Louisiana': 'LA', 'Maine': 'ME',␣
↪'Maryland': 'MD',
    'Massachusetts': 'MA', 'Michigan': 'MI', 'Minnesota': 'MN', 'Mississippi':␣
↪'MS',
    'Missouri': 'MO', 'Montana': 'MT', 'Nebraska': 'NE', 'Nevada': 'NV', 'New␣
↪Hampshire': 'NH',
    'New Jersey': 'NJ', 'New Mexico': 'NM', 'New York': 'NY', 'North Carolina':␣
↪'NC',
    'North Dakota': 'ND', 'Ohio': 'OH', 'Oklahoma': 'OK', 'Oregon': 'OR',␣
↪'Pennsylvania': 'PA',
    'Rhode Island': 'RI', 'South Carolina': 'SC', 'South Dakota': 'SD',␣
↪'Tennessee': 'TN',
    'Texas': 'TX', 'Utah': 'UT', 'Vermont': 'VT', 'Virginia': 'VA',␣
↪'Washington': 'WA',
    'West Virginia': 'WV', 'Wisconsin': 'WI', 'Wyoming': 'WY', 'American Samoa':
↪ 'AS', 'Guam': 'GU', 'Northern Mariana Islands': 'MP',
    'Puerto Rico': 'PR', 'U.S. Virgin Islands': 'VI', 'District of Columbia':␣
↪'DC'}
```

[27]:
```
popu_data['State Abbreviation'] = popu_data['States'].map(state_abbreviations)
```

[28]:
```
popu_data
```

[28]:

	States	Population2020	State Abbreviation
0	California	39,538,223	CA
1	Texas	29,145,505	TX
2	Florida	21,538,187	FL
3	New York	20,201,249	NY
4	Pennsylvania	13,002,700	PA
5	Illinois	12,812,508	IL
6	Ohio	11,799,448	OH
7	Georgia	10,711,908	GA
8	North Carolina	10,439,388	NC
9	Michigan	10,077,331	MI
10	New Jersey	9,288,994	NJ
11	Virginia	8,631,393	VA
12	Washington	7,705,281	WA
13	Arizona	7,151,502	AZ
14	Tennessee	6,910,840	TN
15	Massachusetts	7,029,917	MA
16	Indiana	6,785,528	IN
17	Missouri	6,154,913	MO

18	Maryland	6,177,224	MD
19	Wisconsin	5,893,718	WI
20	Colorado	5,773,714	CO
21	Minnesota	5,706,494	MN
22	South Carolina	5,118,425	SC
23	Alabama	5,024,279	AL
24	Louisiana	4,657,757	LA
25	Kentucky	4,505,836	KY
26	Oregon	4,237,256	OR
27	Oklahoma	3,959,353	OK
28	Connecticut	3,605,944	CT
29	Utah	3,271,616	UT
30	Iowa	3,190,369	IA
31	Puerto Rico	3,285,874	PR
32	Nevada	3,104,614	NV
33	Arkansas	3,011,524	AR
34	Kansas	2,937,880	KS
35	Mississippi	2,961,279	MS
36	New Mexico	2,117,522	NM
37	Nebraska	1,961,504	NE
38	Idaho	1,839,106	ID
39	West Virginia	1,793,716	WV
40	Hawaii	1,455,271	HI
41	New Hampshire	1,377,529	NH
42	Maine	1,362,359	ME
43	Montana	1,084,225	MT
44	Rhode Island	1,097,379	RI
45	Delaware	989,948	DE
46	South Dakota	886,667	SD
47	North Dakota	779,094	ND
48	Alaska	733,391	AK
49	District of Columbia	689,545	DC
50	Vermont	643,077	VT
51	Wyoming	576,851	WY
52	Guam	153,836	GU
53	U.S. Virgin Islands	87,146	VI
54	American Samoa	49,710	AS
55	Northern Mariana Islands	47,329	MP

[29]:
```
# step 4 group by timezone
```

[30]:
```
timezone = {'ON': 'US/Eastern','AK': 'US/Alaska', 'AL': 'US/Central', 'AR': 'US/
↪Central', 'AS': 'US/Samoa',\
            'AZ': 'US/Mountain', 'CA': 'US/Pacific', 'CO': 'US/Mountain', 'CT':␣
↪'US/Eastern', 'DC': 'US/Eastern',\
            'DE': 'US/Eastern', 'FL': 'US/Eastern', 'GA': 'US/Eastern', 'GU':␣
↪'Pacific/Guam', 'HI': 'US/Hawaii',\
```

```
            'IA': 'US/Central', 'ID': 'US/Mountain', 'IL': 'US/Central', 'IN':␣
↪'US/Eastern', 'KS': 'US/Central',\
            'KY': 'US/Eastern', 'LA': 'US/Central', 'MA': 'US/Eastern', 'MD':␣
↪'US/Eastern', 'ME': 'US/Eastern',\
            'MI': 'US/Eastern', 'MN': 'US/Central', 'MO': 'US/Central', 'MP':␣
↪'Pacific/Guam', 'MS': 'US/Central',\
            'MT': 'US/Mountain', 'NC': 'US/Eastern', 'ND': 'US/Central', 'NE':␣
↪'US/Central', 'NH': 'US/Eastern',\
            'NJ': 'US/Eastern', 'NM': 'US/Mountain', 'NV': 'US/Pacific', 'NY':␣
↪'US/Eastern', 'OH': 'US/Eastern',\
            'OK': 'US/Central', 'OR': 'US/Pacific', 'PA': 'US/Eastern', 'PR':␣
↪'America/Puerto_Rico', 'RI': 'US/Eastern',\
            'SC': 'US/Eastern', 'SD': 'US/Central', 'TN': 'US/Central', 'TX':␣
↪'US/Central', 'UT': 'US/Mountain',\
            'VA': 'US/Eastern', 'VI': 'America/Virgin', 'VT': 'US/Eastern',␣
↪'WA': 'US/Pacific', 'WI': 'US/Central',\
            'WV': 'US/Eastern', 'WY': 'US/Mountain', '' : 'US/Pacific', '--':␣
↪'US/Pacific' }
```

[31]:
```
popu_data["Zone"] = popu_data['State Abbreviation']
popu_data['Zone'] = popu_data['Zone'].map(timezone)
```

[32]:
```
popu_data
```

[32]:

	States	Population2020	State Abbreviation	\
0	California	39,538,223	CA	
1	Texas	29,145,505	TX	
2	Florida	21,538,187	FL	
3	New York	20,201,249	NY	
4	Pennsylvania	13,002,700	PA	
5	Illinois	12,812,508	IL	
6	Ohio	11,799,448	OH	
7	Georgia	10,711,908	GA	
8	North Carolina	10,439,388	NC	
9	Michigan	10,077,331	MI	
10	New Jersey	9,288,994	NJ	
11	Virginia	8,631,393	VA	
12	Washington	7,705,281	WA	
13	Arizona	7,151,502	AZ	
14	Tennessee	6,910,840	TN	
15	Massachusetts	7,029,917	MA	
16	Indiana	6,785,528	IN	
17	Missouri	6,154,913	MO	
18	Maryland	6,177,224	MD	
19	Wisconsin	5,893,718	WI	
20	Colorado	5,773,714	CO	
21	Minnesota	5,706,494	MN	

22	South Carolina	5,118,425	SC
23	Alabama	5,024,279	AL
24	Louisiana	4,657,757	LA
25	Kentucky	4,505,836	KY
26	Oregon	4,237,256	OR
27	Oklahoma	3,959,353	OK
28	Connecticut	3,605,944	CT
29	Utah	3,271,616	UT
30	Iowa	3,190,369	IA
31	Puerto Rico	3,285,874	PR
32	Nevada	3,104,614	NV
33	Arkansas	3,011,524	AR
34	Kansas	2,937,880	KS
35	Mississippi	2,961,279	MS
36	New Mexico	2,117,522	NM
37	Nebraska	1,961,504	NE
38	Idaho	1,839,106	ID
39	West Virginia	1,793,716	WV
40	Hawaii	1,455,271	HI
41	New Hampshire	1,377,529	NH
42	Maine	1,362,359	ME
43	Montana	1,084,225	MT
44	Rhode Island	1,097,379	RI
45	Delaware	989,948	DE
46	South Dakota	886,667	SD
47	North Dakota	779,094	ND
48	Alaska	733,391	AK
49	District of Columbia	689,545	DC
50	Vermont	643,077	VT
51	Wyoming	576,851	WY
52	Guam	153,836	GU
53	U.S. Virgin Islands	87,146	VI
54	American Samoa	49,710	AS
55	Northern Mariana Islands	47,329	MP

	Zone
0	US/Pacific
1	US/Central
2	US/Eastern
3	US/Eastern
4	US/Eastern
5	US/Central
6	US/Eastern
7	US/Eastern
8	US/Eastern
9	US/Eastern
10	US/Eastern

```
11           US/Eastern
12           US/Pacific
13          US/Mountain
14           US/Central
15           US/Eastern
16           US/Eastern
17           US/Central
18           US/Eastern
19           US/Central
20          US/Mountain
21           US/Central
22           US/Eastern
23           US/Central
24           US/Central
25           US/Eastern
26           US/Pacific
27           US/Central
28           US/Eastern
29          US/Mountain
30           US/Central
31  America/Puerto_Rico
32           US/Pacific
33           US/Central
34           US/Central
35           US/Central
36          US/Mountain
37           US/Central
38          US/Mountain
39           US/Eastern
40            US/Hawaii
41           US/Eastern
42           US/Eastern
43          US/Mountain
44           US/Eastern
45           US/Eastern
46           US/Central
47           US/Central
48            US/Alaska
49           US/Eastern
50           US/Eastern
51          US/Mountain
52         Pacific/Guam
53       America/Virgin
54             US/Samoa
55         Pacific/Guam
```

[33]:
```
timezone_p = popu_data.groupby(['Zone'])
timezone_p.get_group("US/Eastern")
```

[33]:

	States	Population2020	State Abbreviation	Zone
2	Florida	21,538,187	FL	US/Eastern
3	New York	20,201,249	NY	US/Eastern
4	Pennsylvania	13,002,700	PA	US/Eastern
6	Ohio	11,799,448	OH	US/Eastern
7	Georgia	10,711,908	GA	US/Eastern
8	North Carolina	10,439,388	NC	US/Eastern
9	Michigan	10,077,331	MI	US/Eastern
10	New Jersey	9,288,994	NJ	US/Eastern
11	Virginia	8,631,393	VA	US/Eastern
15	Massachusetts	7,029,917	MA	US/Eastern
16	Indiana	6,785,528	IN	US/Eastern
18	Maryland	6,177,224	MD	US/Eastern
22	South Carolina	5,118,425	SC	US/Eastern
25	Kentucky	4,505,836	KY	US/Eastern
28	Connecticut	3,605,944	CT	US/Eastern
39	West Virginia	1,793,716	WV	US/Eastern
41	New Hampshire	1,377,529	NH	US/Eastern
42	Maine	1,362,359	ME	US/Eastern
44	Rhode Island	1,097,379	RI	US/Eastern
45	Delaware	989,948	DE	US/Eastern
49	District of Columbia	689,545	DC	US/Eastern
50	Vermont	643,077	VT	US/Eastern

[34]:
```
# step 5 drop column and change column index
```

[35]:
```
popu_data1 = popu_data.drop('States', axis=1)
popu_data1
```

[35]:

	Population2020	State Abbreviation	Zone
0	39,538,223	CA	US/Pacific
1	29,145,505	TX	US/Central
2	21,538,187	FL	US/Eastern
3	20,201,249	NY	US/Eastern
4	13,002,700	PA	US/Eastern
5	12,812,508	IL	US/Central
6	11,799,448	OH	US/Eastern
7	10,711,908	GA	US/Eastern
8	10,439,388	NC	US/Eastern
9	10,077,331	MI	US/Eastern
10	9,288,994	NJ	US/Eastern
11	8,631,393	VA	US/Eastern
12	7,705,281	WA	US/Pacific
13	7,151,502	AZ	US/Mountain

14	6,910,840	TN	US/Central
15	7,029,917	MA	US/Eastern
16	6,785,528	IN	US/Eastern
17	6,154,913	MO	US/Central
18	6,177,224	MD	US/Eastern
19	5,893,718	WI	US/Central
20	5,773,714	CO	US/Mountain
21	5,706,494	MN	US/Central
22	5,118,425	SC	US/Eastern
23	5,024,279	AL	US/Central
24	4,657,757	LA	US/Central
25	4,505,836	KY	US/Eastern
26	4,237,256	OR	US/Pacific
27	3,959,353	OK	US/Central
28	3,605,944	CT	US/Eastern
29	3,271,616	UT	US/Mountain
30	3,190,369	IA	US/Central
31	3,285,874	PR	America/Puerto_Rico
32	3,104,614	NV	US/Pacific
33	3,011,524	AR	US/Central
34	2,937,880	KS	US/Central
35	2,961,279	MS	US/Central
36	2,117,522	NM	US/Mountain
37	1,961,504	NE	US/Central
38	1,839,106	ID	US/Mountain
39	1,793,716	WV	US/Eastern
40	1,455,271	HI	US/Hawaii
41	1,377,529	NH	US/Eastern
42	1,362,359	ME	US/Eastern
43	1,084,225	MT	US/Mountain
44	1,097,379	RI	US/Eastern
45	989,948	DE	US/Eastern
46	886,667	SD	US/Central
47	779,094	ND	US/Central
48	733,391	AK	US/Alaska
49	689,545	DC	US/Eastern
50	643,077	VT	US/Eastern
51	576,851	WY	US/Mountain
52	153,836	GU	Pacific/Guam
53	87,146	VI	America/Virgin
54	49,710	AS	US/Samoa
55	47,329	MP	Pacific/Guam

[36]:
```
popu_data1 = popu_data1[['State Abbreviation', 'Zone', 'Population2020']]
```

[37]:
```
popu_data1.rename(columns={'State Abbreviation': 'States'}, inplace=True)
popu_data1
```

[37]:

	States	Zone	Population2020
0	CA	US/Pacific	39,538,223
1	TX	US/Central	29,145,505
2	FL	US/Eastern	21,538,187
3	NY	US/Eastern	20,201,249
4	PA	US/Eastern	13,002,700
5	IL	US/Central	12,812,508
6	OH	US/Eastern	11,799,448
7	GA	US/Eastern	10,711,908
8	NC	US/Eastern	10,439,388
9	MI	US/Eastern	10,077,331
10	NJ	US/Eastern	9,288,994
11	VA	US/Eastern	8,631,393
12	WA	US/Pacific	7,705,281
13	AZ	US/Mountain	7,151,502
14	TN	US/Central	6,910,840
15	MA	US/Eastern	7,029,917
16	IN	US/Eastern	6,785,528
17	MO	US/Central	6,154,913
18	MD	US/Eastern	6,177,224
19	WI	US/Central	5,893,718
20	CO	US/Mountain	5,773,714
21	MN	US/Central	5,706,494
22	SC	US/Eastern	5,118,425
23	AL	US/Central	5,024,279
24	LA	US/Central	4,657,757
25	KY	US/Eastern	4,505,836
26	OR	US/Pacific	4,237,256
27	OK	US/Central	3,959,353
28	CT	US/Eastern	3,605,944
29	UT	US/Mountain	3,271,616
30	IA	US/Central	3,190,369
31	PR	America/Puerto_Rico	3,285,874
32	NV	US/Pacific	3,104,614
33	AR	US/Central	3,011,524
34	KS	US/Central	2,937,880
35	MS	US/Central	2,961,279
36	NM	US/Mountain	2,117,522
37	NE	US/Central	1,961,504
38	ID	US/Mountain	1,839,106
39	WV	US/Eastern	1,793,716
40	HI	US/Hawaii	1,455,271
41	NH	US/Eastern	1,377,529
42	ME	US/Eastern	1,362,359
43	MT	US/Mountain	1,084,225
44	RI	US/Eastern	1,097,379
45	DE	US/Eastern	989,948

46	SD	US/Central	886,667
47	ND	US/Central	779,094
48	AK	US/Alaska	733,391
49	DC	US/Eastern	689,545
50	VT	US/Eastern	643,077
51	WY	US/Mountain	576,851
52	GU	Pacific/Guam	153,836
53	VI	America/Virgin	87,146
54	AS	US/Samoa	49,710
55	MP	Pacific/Guam	47,329

The population data we adopted here is "List of U.S. states and territories by population". It has the State and territory rankings by populations. This datasets has provided the census population of 2020, and the predition of the population of 2023, as well as the change from 2010-2020. However, for my project I only need the population of 2020 for each states. First the scraped the data from this wikipedia page using beautiful soup, and then made some adjustments for the following transformation of my project. I changed the state names in full to abbreviation in order to align with the csv file I used for the covid 19 vaccination information. Besides I also grouped the states by timezones, to prepare for seeing the vaccination in each different zones.

Since this datasets primarily provides statistical data regarding the population of U.S. states and territories. It provides neutral record and generally does not have ethical implications. A couple of potential ethical implications that I can think of are: 1. Analyze and further publish this data along with vaccination data may unveil, to some extent, influence the public opinions and impact people's atitude about vaccination. 2. The data can potentials demonstrates the disparity between different states, especially if used to support biased or discriminatory policies. Misinterpretation can occur if the data is taken out of context or simplified too much.

[38]:
```
file_path = "popu_data1.csv"
popu_data1.to_csv(file_path, index=False)
```

[ ]:
```
```