Milestone 3

July 16, 2023

[39]:

```
# Import libraries
import pandas as pd

# Establish URL
url = 'https://en.wikipedia.org/wiki/List_of_U.S.
 ↪_states_and_territories_by_GDP#GDP_by_state'

# Read the website and obtain tables
tables = pd.read_html(url)
```

[40]:

```
# Creates a Dataframes using the first table
wiki_data = pd.DataFrame(tables[0])

wiki_data.head()
```

[40]:

```
  State or federal district  \
  State or federal district
0                       NaN
1               California *
2                    Texas *
3                 New York *
4                  Florida *

  Nominal GDPat current prices 2022(millions ofU.S. dollars)[1]              \
                                                              2022        2023
0                                                        NaN           NaN
1                                                  3598103.0     3755487.0
2                                                  2355960.0     2436346.0
3                                                  2053180.0     2135672.0
4                                                  1389070.0     1468015.0

  Annual GDP changeat current prices 2022(21-22)[1]  \
  Annual GDP changeat current prices 2022(21-22)[1]
0                                               NaN
1                                          224862.0
2                                          304191.0
3                                          151883.0
```

```
4                                          133482.0

                                                       \
  Annual GDP changeat current prices 2022(21-22)[1].1
0                                                 NaN
1                                               11.6%
2                                               32.6%
3                                                7.5%
4                                               10.9%

  Real GDP growthrate (2021-2022)[1] Nominal GDP per capita 2022[1][3]  \
  Real GDP growthrate (2021-2022)[1]                              2022
0                                NaN                               NaN
1                               7.8%                           $92,190
2                               5.6%                           $78,456
3                               5.0%                          $104,344
4                               6.9%                           $62,446

           % of national[1]
      2021             2022    2021
0      NaN              NaN     NaN
1  $85,316           14.69%  14.49%
2  $70,398            8.69%   8.55%
3  $94,118            8.11%   8.31%
4  $58,295            5.37%   5.34%
```

```
[41]: wiki_data.columns
```

```
[41]: MultiIndex([(                                     'State or federal district',
…),
            ('Nominal GDPat current prices 2022(millions ofU.S. dollars)[1]',
…),
            ('Nominal GDPat current prices 2022(millions ofU.S. dollars)[1]',
…),
            (              'Annual GDP changeat current prices 2022(21-22)[1]',
…),
            (              'Annual GDP changeat current prices 2022(21-22)[1]',
…),
            (                            'Real GDP growthrate (2021-2022)[1]',
…),
            (                             'Nominal GDP per capita 2022[1][3]',
…),
            (                             'Nominal GDP per capita 2022[1][3]',
…),
            (                                              '% of national[1]',
…),
            (                                              '% of national[1]',
…)],
            )
```

Step 1

```
[42]: """
As seen above the columns are in a multilevel index. Joining the data to the␣
 ↪other data sets will be easier
if they have the same amount of levels.
To do that, will use the DropLevel, and index the first level to remove it.
"""
# Drops first level
wiki_data.columns = wiki_data.columns.droplevel(0)
wiki_data.head()
```

```
[42]:   State or federal district       2022       2023  \
0                       NaN        NaN        NaN
1              California *  3598103.0  3755487.0
2                   Texas *  2355960.0  2436346.0
3                New York *  2053180.0  2135672.0
4                 Florida *  1389070.0  1468015.0

   Annual GDP changeat current prices 2022(21-22)[1]  \
0                                                NaN
1                                           224862.0
2                                           304191.0
3                                           151883.0
4                                           133482.0

  Annual GDP changeat current prices 2022(21-22)[1].1  \
0                                                NaN
1                                              11.6%
2                                              32.6%
3                                               7.5%
4                                              10.9%

  Real GDP growthrate (2021-2022)[1]      2022     2021     2022     2021
0                                NaN       NaN      NaN     NaN     NaN
1                               7.8%   $92,190  $85,316  14.69%  14.49%
2                               5.6%   $78,456  $70,398   8.69%   8.55%
3                               5.0%  $104,344  $94,118   8.11%   8.31%
4                               6.9%   $62,446  $58,295   5.37%   5.34%
```

```
[43]: """
With that, the columns are all in one level and we can rename the columns as␣
 ↪needed.
"""
```

```
wiki_data.columns
```

```
[43]: Index(['State or federal district', '2022', '2023',
              'Annual GDP changeat current prices 2022(21-22)[1]',
              'Annual GDP changeat current prices 2022(21-22)[1].1',
              'Real GDP growthrate (2021-2022)[1]', '2022', '2021', '2022', '2021'],
             dtype='object')
```

Step 2

```
[44]: """
      Now that we have removed the top index layer, the first column is all NaN␣
       ↪values.
      We can use the dropna function to drop all nulls from the data.
      """
      # Removes any NaN values
      wiki_data = wiki_data.dropna()
      wiki_data.head()
```

```
[44]:   State or federal district       2022       2023  \
      1              California *  3598103.0  3755487.0
      2                   Texas *  2355960.0  2436346.0
      3                New York *  2053180.0  2135672.0
      4                 Florida *  1389070.0  1468015.0
      5                Illinois *  1033310.0  1071552.0

         Annual GDP changeat current prices 2022(21-22)[1]  \
      1                                           224862.0
      2                                           304191.0
      3                                           151883.0
      4                                           133482.0
      5                                            87636.0

        Annual GDP changeat current prices 2022(21-22)[1].1  \
      1                                              11.6%
      2                                              32.6%
      3                                               7.5%
      4                                              10.9%
      5                                               9.3%

        Real GDP growthrate (2021-2022)[1]      2022     2021    2022    2021
      1                               7.8%   $92,190  $85,316  14.69%  14.49%
      2                               5.6%   $78,456  $70,398   8.69%   8.55%
      3                               5.0%  $104,344  $94,118   8.11%   8.31%
      4                               6.9%   $62,446  $58,295   5.37%   5.34%
      5                               5.0%   $82,126  $73,811   4.11%   4.13%
```

Step 3

[45]:

```
"""
The last 2 columns of this data provide % of national which will not be needed␣
 ↪for this analysis.
These two columns can be dropped using iloc.
"""
# Removes the last 2 columns
wiki_data = wiki_data.iloc[:, :-2]
wiki_data.head()
```

[45]:

```
  State or federal district       2022       2023  \
1              California *  3598103.0  3755487.0
2                   Texas *  2355960.0  2436346.0
3                New York *  2053180.0  2135672.0
4                 Florida *  1389070.0  1468015.0
5                Illinois *  1033310.0  1071552.0

   Annual GDP changeat current prices 2022(21-22)[1]  \
1                                           224862.0
2                                           304191.0
3                                           151883.0
4                                           133482.0
5                                            87636.0

  Annual GDP changeat current prices 2022(21-22)[1].1  \
1                                              11.6%
2                                              32.6%
3                                               7.5%
4                                              10.9%
5                                               9.3%

  Real GDP growthrate (2021-2022)[1]      2022     2021
1                               7.8%   $92,190  $85,316
2                               5.6%   $78,456  $70,398
3                               5.0%  $104,344  $94,118
4                               6.9%   $62,446  $58,295
5                               5.0%   $82,126  $73,811
```

Step 4

[46]:

```
"""
Since the top index level was dropped in a previous we need to rename the last␣
 ↪two columns to be 'Per Capita 2022' and 'Per Capita 2021'.
This can be completed with the rename funciton in pandas.
"""
# Renames 2022 and 2021
wiki_data = wiki_data.rename(columns={'2022': 'PerCapitaGDP_2022', '2021':␣
 ↪'PerCapitaGDP_2021'})
```

```
wiki_data.head()
```

[46]:

```
  State or federal district PerCapitaGDP_2022       2023  \
1               California *         3598103.0  3755487.0
2                    Texas *         2355960.0  2436346.0
3                 New York *         2053180.0  2135672.0
4                  Florida *         1389070.0  1468015.0
5                 Illinois *         1033310.0  1071552.0

   Annual GDP changeat current prices 2022(21-22)[1]  \
1                                           224862.0
2                                           304191.0
3                                           151883.0
4                                           133482.0
5                                            87636.0

  Annual GDP changeat current prices 2022(21-22)[1].1  \
1                                              11.6%
2                                              32.6%
3                                               7.5%
4                                              10.9%
5                                               9.3%

  Real GDP growthrate (2021-2022)[1] PerCapitaGDP_2022 PerCapitaGDP_2021
1                              7.8%           $92,190           $85,316
2                              5.6%           $78,456           $70,398
3                              5.0%          $104,344           $94,118
4                              6.9%           $62,446           $58,295
5                              5.0%           $82,126           $73,811
```

Step 5

[47]:

```
"""
There were two columns with the same name. The column in position 2 needs to be␣
 ↪renamed to GDP 2022.
This will done using column indexing so that the other column with the same␣
 ↪name does not get changed again.
"""
# Indexes second column and changes name
wiki_data.columns.values[1] = 'GDP_2022'
wiki_data.head()
```

[47]:

```
  State or federal district   GDP_2022       2023  \
1               California *  3598103.0  3755487.0
2                    Texas *  2355960.0  2436346.0
3                 New York *  2053180.0  2135672.0
4                  Florida *  1389070.0  1468015.0
```

```
5                Illinois *  1033310.0  1071552.0

   Annual GDP changeat current prices 2022(21-22)[1]  \
1                                           224862.0
2                                           304191.0
3                                           151883.0
4                                           133482.0
5                                            87636.0

  Annual GDP changeat current prices 2022(21-22)[1].1  \
1                                              11.6%
2                                              32.6%
3                                               7.5%
4                                              10.9%
5                                               9.3%

  Real GDP growthrate (2021-2022)[1] PerCapitaGDP_2022 PerCapitaGDP_2021
1                               7.8%           $92,190           $85,316
2                               5.6%           $78,456           $70,398
3                               5.0%          $104,344           $94,118
4                               6.9%           $62,446           $58,295
5                               5.0%           $82,126           $73,811
```

Step 5

[48]:

```python
"""
Next, the third column, 2023, had another level above it which stated it was GDP for 2023.
This column will need to be renamed to display the accurate information.
"""
# Renames 2023 Columm
wiki_data = wiki_data.rename(columns={'2023': 'GDP_2023'})
wiki_data.head()
```

[48]:

```
  State or federal district   GDP_2022   GDP_2023  \
1             California *  3598103.0  3755487.0
2                  Texas *  2355960.0  2436346.0
3               New York *  2053180.0  2135672.0
4                Florida *  1389070.0  1468015.0
5               Illinois *  1033310.0  1071552.0

   Annual GDP changeat current prices 2022(21-22)[1]  \
1                                           224862.0
2                                           304191.0
3                                           151883.0
4                                           133482.0
5                                            87636.0
```

```
  Annual GDP changeat current prices 2022(21-22)[1].1  \
1                                               11.6%
2                                               32.6%
3                                                7.5%
4                                               10.9%
5                                                9.3%

  Real GDP growthrate (2021-2022)[1] PerCapitaGDP_2022 PerCapitaGDP_2021
1                               7.8%           $92,190           $85,316
2                               5.6%           $78,456           $70,398
3                               5.0%          $104,344           $94,118
4                               6.9%           $62,446           $58,295
5                               5.0%           $82,126           $73,811
```

Step 6

[49]:

```
"""
The last two columns have dollar signs and the others do not.
To keep this consistent with the current table and the tables from the other␣
 ↪datasets the $ and , will be removed.
This can be completed with str.replace function.
"""
# Removes $ from string
wiki_data['PerCapitaGDP_2022'] = wiki_data['PerCapitaGDP_2022'].str.
 ↪replace('$', '')
wiki_data['PerCapitaGDP_2021'] = wiki_data['PerCapitaGDP_2021'].str.
 ↪replace('$', '')
# Removes , from string
wiki_data['PerCapitaGDP_2022'] = wiki_data['PerCapitaGDP_2022'].str.
 ↪replace(',', '')
wiki_data['PerCapitaGDP_2021'] = wiki_data['PerCapitaGDP_2021'].str.
 ↪replace(',', '')
wiki_data.head()
```

```
/var/folders/sr/xvmzsbj91c91yq0f0qnq71xh0000gn/T/ipykernel_1925/3912393484.py:6:
FutureWarning: The default value of regex will change from True to False in a
future version. In addition, single character regular expressions will *not* be
treated as literal strings when regex=True.
  wiki_data['PerCapitaGDP_2022'] =
wiki_data['PerCapitaGDP_2022'].str.replace('$', '')
/var/folders/sr/xvmzsbj91c91yq0f0qnq71xh0000gn/T/ipykernel_1925/3912393484.py:7:
FutureWarning: The default value of regex will change from True to False in a
future version. In addition, single character regular expressions will *not* be
treated as literal strings when regex=True.
  wiki_data['PerCapitaGDP_2021'] =
wiki_data['PerCapitaGDP_2021'].str.replace('$', '')
```

[49]:
```
   State or federal district   GDP_2022   GDP_2023  \
1               California *  3598103.0  3755487.0
2                    Texas *  2355960.0  2436346.0
3                 New York *  2053180.0  2135672.0
4                  Florida *  1389070.0  1468015.0
5                 Illinois *  1033310.0  1071552.0

    Annual GDP changeat current prices 2022(21-22)[1]  \
1                                           224862.0
2                                           304191.0
3                                           151883.0
4                                           133482.0
5                                            87636.0

   Annual GDP changeat current prices 2022(21-22)[1].1  \
1                                              11.6%
2                                              32.6%
3                                               7.5%
4                                              10.9%
5                                               9.3%

   Real GDP growthrate (2021-2022)[1] PerCapitaGDP_2022 PerCapitaGDP_2021
1                               7.8%             92190             85316
2                               5.6%             78456             70398
3                               5.0%            104344             94118
4                               6.9%             62446             58295
5                               5.0%             82126             73811
```

Step 7

[50]:
```
"""
The middle three columns provide information that will not be used in this␣
 ↪study and can be ommitted.
They can be removed using column indexing.
"""
# Drops middle three columns
wiki_data = wiki_data.drop(wiki_data.columns[3:6], axis=1)
wiki_data.head()
```

[50]:
```
   State or federal district   GDP_2022   GDP_2023 PerCapitaGDP_2022  \
1               California *  3598103.0  3755487.0             92190
2                    Texas *  2355960.0  2436346.0             78456
3                 New York *  2053180.0  2135672.0            104344
4                  Florida *  1389070.0  1468015.0             62446
5                 Illinois *  1033310.0  1071552.0             82126

   PerCapitaGDP_2021
```

```
1             85316
2             70398
3             94118
4             58295
5             73811
```

Step 8

[51]:

```
"""
To keep the data consistent, we will convert all the numeric columns into␣
 ↪floats.
This can be done by using as type and converting the strings into floats.
"""
# Converts last two columns into floats
wiki_data['PerCapitaGDP_2022'] = wiki_data['PerCapitaGDP_2022'].astype(float)
wiki_data['PerCapitaGDP_2021'] = wiki_data['PerCapitaGDP_2021'].astype(float)
wiki_data.head()
```

[51]:

```
  State or federal district   GDP_2022   GDP_2023  PerCapitaGDP_2022  \
1             California *  3598103.0  3755487.0            92190.0
2                  Texas *  2355960.0  2436346.0            78456.0
3               New York *  2053180.0  2135672.0           104344.0
4                Florida *  1389070.0  1468015.0            62446.0
5               Illinois *  1033310.0  1071552.0            82126.0

   PerCapitaGDP_2021
1            85316.0
2            70398.0
3            94118.0
4            58295.0
5            73811.0
```

[52]:

```
wiki_data.dtypes
```

[52]:

```
State or federal district     object
GDP_2022                     float64
GDP_2023                     float64
PerCapitaGDP_2022            float64
PerCapitaGDP_2021            float64
dtype: object
```

Step 9

[53]:

```
"""
For easier reading, we will add commas to each numeric value.
This will be done by using applymap and applying this to all the columns after␣
 ↪the first.
"""
```

```
# Adds comma separators to last 4 columns
wiki_data.loc[:,1:] = wiki_data.iloc[:,1:].applymap(lambda x: '{:,}'.format(x))
wiki_data.head()
```

```
/var/folders/sr/xvmzsbj91c91yq0f0qnq71xh0000gn/T/ipykernel_1925/4136283864.py:1:
FutureWarning: Slicing a positional slice with .loc is not supported, and will
raise TypeError in a future version.  Use .loc with labels or .iloc with
positions instead.
  wiki_data.loc[:,1:] = wiki_data.iloc[:,1:].applymap(lambda x:
'{:,}'.format(x))
```

```
[53]:   State or federal district     GDP_2022     GDP_2023 PerCapitaGDP_2022  \
      1               California *  3,598,103.0  3,755,487.0          92,190.0
      2                    Texas *  2,355,960.0  2,436,346.0          78,456.0
      3                 New York *  2,053,180.0  2,135,672.0         104,344.0
      4                  Florida *  1,389,070.0  1,468,015.0          62,446.0
      5                 Illinois *  1,033,310.0  1,071,552.0          82,126.0

        PerCapitaGDP_2021
      1          85,316.0
      2          70,398.0
      3          94,118.0
      4          58,295.0
      5          73,811.0
```

Step 10

```
[55]: """
      The next item that will be changed will be the names of the states.
      Since some of them contain asterisks, it will not be possible to join them to␣
       ↪another dataset.
      This can be completed with str.replace function.
      """
      # Removes asterisk from State or federal district column
      wiki_data['State or federal district'] = wiki_data['State or federal district'].
       ↪str.replace('*', '')
      wiki_data.head()
```

```
/var/folders/sr/xvmzsbj91c91yq0f0qnq71xh0000gn/T/ipykernel_1925/3891370520.py:6:
FutureWarning: The default value of regex will change from True to False in a
future version. In addition, single character regular expressions will *not* be
treated as literal strings when regex=True.
  wiki_data['State or federal district'] = wiki_data['State or federal
district'].str.replace('*', '')
```

```
[55]:   State or federal district     GDP_2022     GDP_2023 PerCapitaGDP_2022  \
      1                California  3,598,103.0  3,755,487.0          92,190.0
      2                     Texas  2,355,960.0  2,436,346.0          78,456.0
```

```
3                 New York  2,053,180.0  2,135,672.0         104,344.0
4                  Florida  1,389,070.0  1,468,015.0          62,446.0
5                 Illinois  1,033,310.0  1,071,552.0          82,126.0

  PerCapitaGDP_2021
1         85,316.0
2         70,398.0
3         94,118.0
4         58,295.0
5         73,811.0
```

Step 11

[56]:

```
"""
The last item that will be changed is the name of the first column.
When joining, the same column for state will be needed.
"""
# Renames State or federal district column
wiki_data = wiki_data.rename(columns={'State or federal district': 'StateName'})
wiki_data.head()
```

[56]:

```
    StateName     GDP_2022     GDP_2023 PerCapitaGDP_2022 PerCapitaGDP_2021
1  California  3,598,103.0  3,755,487.0          92,190.0          85,316.0
2       Texas  2,355,960.0  2,436,346.0          78,456.0          70,398.0
3    New York  2,053,180.0  2,135,672.0         104,344.0          94,118.0
4     Florida  1,389,070.0  1,468,015.0          62,446.0          58,295.0
5    Illinois  1,033,310.0  1,071,552.0          82,126.0          73,811.0
```

With this data, the ethical implications that could be found can involve the use of wide scale data versus smaller scale. Since GDP by city can be hard to be obtain, the state GDP can be used. Howerver, this use can give the average but will lack the granularity city GDP can offer.