house -Copy2

November 24, 2024

1 Simple Webscraping Example with Beautiful Soup

[18]:
```
# Import all neccessary libraries

from bs4 import BeautifulSoup
import urllib.request
import pandas as pd
```

[19]:
```
# Assign the URL to a variable

url = "https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/details/
 ↪page.cfm?
 ↪Lang=E&SearchText=M1C&DGUIDlist=2021A0011M1C&GENDERlist=1,2,3&STATISTIClist=1,4&HEADERlist=



# use the urlopen function to open the webpage
html = urllib.request.urlopen(url)

# show object html
html
```

[19]:
```
<http.client.HTTPResponse at 0x7f245ebdeb90>
```

[20]:
```
# Create the BeautifulSoup object

html_to_parse = BeautifulSoup(html, "html.parser")
```

[21]:
```
# create a list of tables.  There is only 1 table in this webpage

tables = html_to_parse.find_all("table")
print(f"Number of tables found: {len(tables)}")
```

```
Number of tables found: 1
```

[22]:
```
# Create list of all the <th> tags in the table that has the title␣
 ↪"2021A0011M1C - Population, 2021 - Counts - Total"
```

```
td = tables[0].find(attrs={"title":"2021A0011M1C - 40 to 44 years - Counts -␣
↪Total"})
```

[23]:
```
td
```

[23]:
```
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M1C
geo2021A0011M1Cstat1 geo2021A0011M1Cstat1gen1" title="2021A0011M1C - 40 to 44
years - Counts - Total">  2,090</td>
```

[24]:
```
# convert to float
float(td.text.replace(",", ""))
```

[24]:
```
2090.0
```

1.1 Create a script that will look up from a list of Postal codes

[25]:
```
import urllib.parse as urlparse
from urllib.parse import urlencode
```

[26]:
```
# A list of postal code from the previous part

postal = ['M3A', 'M4A', 'M5A', 'M6A', 'M7A','M5E', 'M4E', 'M6E', 'M5G',␣
↪'M6G','M2H','M3H','M4H','M5H','M6H',
'M1J','M2J', 'M3J',
       'M4J',
       'M5J',
       'M5K',
       'M6K',
       'M1L',
       'M2L',
       'M3L',
       'M4L',
       'M5L',
       'M6L',
       'M9L',
       'M1M',
       'M2M',
       'M3M',
       'M4M',
       'M5M',
       'M6M',
       'M9M',
       'M1N',
       'M2N',
       'M3N',
       'M4N',
       'M5N',
```

```
        'M6N',  'M9N',
        'M1P',
        'M2P',
        'M4P',
        'M5P',
        'M6P',
        'M9P',
        'M1R',
        'M2R',
        'M4R',
        'M5R',
        'M6R',
        'M7R',
        'M9R',
        'M1S',  'M4S',
        'M5S',
        'M6S',
        'M1T',
        'M4T',
        'M5T',
        'M1V',
        'M4V',
        'M5V',
        'M8V',
        'M9V',
        'M1W',
        'M4W',
        'M5W',
        'M8W',
        'M9W',
        'M1X',
        'M4X',
        'M5X',
        'M8X',
        'M7Y',
        'M8Y',
        'M8Z',]
```

[27]:
```
# Creating Empty DataFrame and Storing it in variable df

df = pd.DataFrame(columns = ['postal_code', 'data', 'value'])
```

[ ]:
```
# Loop through each postal code
import numpy as np

for i in postal:
```

```
    url = "https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/
↪details/page.cfm?Lang=E"
    params = {
        'SearchText': i,
        'DGUIDlist': '2021A0011'+i
    }

    # this part switches up the postal code parameter in the url
    url_parts = list(urlparse.urlparse(url))
    query = dict(urlparse.parse_qsl(url_parts[4]))
    query.update(params)

    url_parts[4] = urlencode(query)
    query = urlparse.urlunparse(url_parts)

    # the following code is similar to the above
    html = urllib.request.urlopen(query)
    html_to_parse = BeautifulSoup(html, "html.parser")
    tables = html_to_parse.find_all("table")
    print(f"Number of tables found: {len(tables)}")


    tables = html_to_parse.find_all("table")
    if len(tables) == 0:
        print("No tables found on this page.")
        continue  # Skip to the next iteration




    # change the title to find the data you want
    title = (f"2021A0011{i} - 40 to 44 years - Counts - Total")
    td = tables[0].find(attrs={"title":title})
    print(td)


    if td:
        try:
            # Try to convert text to float
            value = float(td.text.replace(",", ""))
        except ValueError:
            # Handle suppressed data or invalid values
            value = np.nan
            print(f"Data suppressed or invalid for {i}, setting value to NaN.")

        df.loc[len(df.index)] = [i, title, value]
    else:
```

```
        print(f"No data found for title: {title}")
```

```
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M3A
geo2021A0011M3Astat1 geo2021A0011M3Astat1gen1" title="2021A0011M3A - 40 to 44
years - Counts - Total">  2,340</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M4A
geo2021A0011M4Astat1 geo2021A0011M4Astat1gen1" title="2021A0011M4A - 40 to 44
years - Counts - Total">  1,020</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M5A
geo2021A0011M5Astat1 geo2021A0011M5Astat1gen1" title="2021A0011M5A - 40 to 44
years - Counts - Total">  3,570</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M6A
geo2021A0011M6Astat1 geo2021A0011M6Astat1gen1" title="2021A0011M6A - 40 to 44
years - Counts - Total">  1,585</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M7A
geo2021A0011M7Astat1 geo2021A0011M7Astat1gen1" title="2021A0011M7A - 40 to 44
years - Counts - Total"><abbr title="suppressed to meet the confidentiality
requirements of the Statistics Act">x</abbr></td>
Data suppressed or invalid for M7A, setting value to NaN.
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M5E
geo2021A0011M5Estat1 geo2021A0011M5Estat1gen1" title="2021A0011M5E - 40 to 44
years - Counts - Total">   715</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M4E
geo2021A0011M4Estat1 geo2021A0011M4Estat1gen1" title="2021A0011M4E - 40 to 44
years - Counts - Total">  2,055</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M6E
geo2021A0011M6Estat1 geo2021A0011M6Estat1gen1" title="2021A0011M6E - 40 to 44
years - Counts - Total">  2,765</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M5G
geo2021A0011M5Gstat1 geo2021A0011M5Gstat1gen1" title="2021A0011M5G - 40 to 44
years - Counts - Total">   490</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M6G
geo2021A0011M6Gstat1 geo2021A0011M6Gstat1gen1" title="2021A0011M6G - 40 to 44
years - Counts - Total">  1,995</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M2H
geo2021A0011M2Hstat1 geo2021A0011M2Hstat1gen1" title="2021A0011M2H - 40 to 44
```

```
years - Counts - Total">  1,185</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M3H
geo2021A0011M3Hstat1 geo2021A0011M3Hstat1gen1" title="2021A0011M3H - 40 to 44
years - Counts - Total">  2,915</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M4H
geo2021A0011M4Hstat1 geo2021A0011M4Hstat1gen1" title="2021A0011M4H - 40 to 44
years - Counts - Total">  1,180</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M5H
geo2021A0011M5Hstat1 geo2021A0011M5Hstat1gen1" title="2021A0011M5H - 40 to 44
years - Counts - Total">   170</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M6H
geo2021A0011M6Hstat1 geo2021A0011M6Hstat1gen1" title="2021A0011M6H - 40 to 44
years - Counts - Total">  3,495</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M1J
geo2021A0011M1Jstat1 geo2021A0011M1Jstat1gen1" title="2021A0011M1J - 40 to 44
years - Counts - Total">  2,360</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M2J
geo2021A0011M2Jstat1 geo2021A0011M2Jstat1gen1" title="2021A0011M2J - 40 to 44
years - Counts - Total">  4,075</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M3J
geo2021A0011M3Jstat1 geo2021A0011M3Jstat1gen1" title="2021A0011M3J - 40 to 44
years - Counts - Total">  1,615</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M4J
geo2021A0011M4Jstat1 geo2021A0011M4Jstat1gen1" title="2021A0011M4J - 40 to 44
years - Counts - Total">  2,965</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M5J
geo2021A0011M5Jstat1 geo2021A0011M5Jstat1gen1" title="2021A0011M5J - 40 to 44
years - Counts - Total">  1,050</td>
Number of tables found: 0
No tables found on this page.
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M6K
geo2021A0011M6Kstat1 geo2021A0011M6Kstat1gen1" title="2021A0011M6K - 40 to 44
years - Counts - Total">  3,415</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M1L
geo2021A0011M1Lstat1 geo2021A0011M1Lstat1gen1" title="2021A0011M1L - 40 to 44
years - Counts - Total">  2,510</td>
Number of tables found: 1
```

```
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M2L
geo2021A0011M2Lstat1 geo2021A0011M2Lstat1gen1" title="2021A0011M2L - 40 to 44
years - Counts - Total">   585</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M3L
geo2021A0011M3Lstat1 geo2021A0011M3Lstat1gen1" title="2021A0011M3L - 40 to 44
years - Counts - Total">  1,340</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M4L
geo2021A0011M4Lstat1 geo2021A0011M4Lstat1gen1" title="2021A0011M4L - 40 to 44
years - Counts - Total">  2,790</td>
Number of tables found: 0
No tables found on this page.
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M6L
geo2021A0011M6Lstat1 geo2021A0011M6Lstat1gen1" title="2021A0011M6L - 40 to 44
years - Counts - Total">  1,295</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M9L
geo2021A0011M9Lstat1 geo2021A0011M9Lstat1gen1" title="2021A0011M9L - 40 to 44
years - Counts - Total">   670</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M1M
geo2021A0011M1Mstat1 geo2021A0011M1Mstat1gen1" title="2021A0011M1M - 40 to 44
years - Counts - Total">  1,295</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M2M
geo2021A0011M2Mstat1 geo2021A0011M2Mstat1gen1" title="2021A0011M2M - 40 to 44
years - Counts - Total">  1,900</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M3M
geo2021A0011M3Mstat1 geo2021A0011M3Mstat1gen1" title="2021A0011M3M - 40 to 44
years - Counts - Total">  1,805</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M4M
geo2021A0011M4Mstat1 geo2021A0011M4Mstat1gen1" title="2021A0011M4M - 40 to 44
years - Counts - Total">  2,175</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M5M
geo2021A0011M5Mstat1 geo2021A0011M5Mstat1gen1" title="2021A0011M5M - 40 to 44
years - Counts - Total">  1,760</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M6M
geo2021A0011M6Mstat1 geo2021A0011M6Mstat1gen1" title="2021A0011M6M - 40 to 44
years - Counts - Total">  2,860</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M9M
geo2021A0011M9Mstat1 geo2021A0011M9Mstat1gen1" title="2021A0011M9M - 40 to 44
```

```
years - Counts - Total">  1,660</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M1N
geo2021A0011M1Nstat1 geo2021A0011M1Nstat1gen1" title="2021A0011M1N - 40 to 44
years - Counts - Total">  1,600</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M2N
geo2021A0011M2Nstat1 geo2021A0011M2Nstat1gen1" title="2021A0011M2N - 40 to 44
years - Counts - Total">  5,090</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M3N
geo2021A0011M3Nstat1 geo2021A0011M3Nstat1gen1" title="2021A0011M3N - 40 to 44
years - Counts - Total">  2,450</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M4N
geo2021A0011M4Nstat1 geo2021A0011M4Nstat1gen1" title="2021A0011M4N - 40 to 44
years - Counts - Total">   930</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M5N
geo2021A0011M5Nstat1 geo2021A0011M5Nstat1gen1" title="2021A0011M5N - 40 to 44
years - Counts - Total">  1,070</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M6N
geo2021A0011M6Nstat1 geo2021A0011M6Nstat1gen1" title="2021A0011M6N - 40 to 44
years - Counts - Total">  2,905</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M9N
geo2021A0011M9Nstat1 geo2021A0011M9Nstat1gen1" title="2021A0011M9N - 40 to 44
years - Counts - Total">  1,765</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M1P
geo2021A0011M1Pstat1 geo2021A0011M1Pstat1gen1" title="2021A0011M1P - 40 to 44
years - Counts - Total">  2,615</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M2P
geo2021A0011M2Pstat1 geo2021A0011M2Pstat1gen1" title="2021A0011M2P - 40 to 44
years - Counts - Total">   445</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M4P
geo2021A0011M4Pstat1 geo2021A0011M4Pstat1gen1" title="2021A0011M4P - 40 to 44
years - Counts - Total">  1,885</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M5P
geo2021A0011M5Pstat1 geo2021A0011M5Pstat1gen1" title="2021A0011M5P - 40 to 44
years - Counts - Total">  1,165</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M6P
geo2021A0011M6Pstat1 geo2021A0011M6Pstat1gen1" title="2021A0011M6P - 40 to 44
```

```
years - Counts - Total">  3,360</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M9P
geo2021A0011M9Pstat1 geo2021A0011M9Pstat1gen1" title="2021A0011M9P - 40 to 44
years - Counts - Total">  1,250</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M1R
geo2021A0011M1Rstat1 geo2021A0011M1Rstat1gen1" title="2021A0011M1R - 40 to 44
years - Counts - Total">  1,880</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M2R
geo2021A0011M2Rstat1 geo2021A0011M2Rstat1gen1" title="2021A0011M2R - 40 to 44
years - Counts - Total">  2,755</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M4R
geo2021A0011M4Rstat1 geo2021A0011M4Rstat1gen1" title="2021A0011M4R - 40 to 44
years - Counts - Total">   820</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M5R
geo2021A0011M5Rstat1 geo2021A0011M5Rstat1gen1" title="2021A0011M5R - 40 to 44
years - Counts - Total">  1,475</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M6R
geo2021A0011M6Rstat1 geo2021A0011M6Rstat1gen1" title="2021A0011M6R - 40 to 44
years - Counts - Total">  1,595</td>
Number of tables found: 0
No tables found on this page.
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M9R
geo2021A0011M9Rstat1 geo2021A0011M9Rstat1gen1" title="2021A0011M9R - 40 to 44
years - Counts - Total">  2,065</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M1S
geo2021A0011M1Sstat1 geo2021A0011M1Sstat1gen1" title="2021A0011M1S - 40 to 44
years - Counts - Total">  1,915</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M4S
geo2021A0011M4Sstat1 geo2021A0011M4Sstat1gen1" title="2021A0011M4S - 40 to 44
years - Counts - Total">  2,315</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M5S
geo2021A0011M5Sstat1 geo2021A0011M5Sstat1gen1" title="2021A0011M5S - 40 to 44
years - Counts - Total">   855</td>
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M6S
geo2021A0011M6Sstat1 geo2021A0011M6Sstat1gen1" title="2021A0011M6S - 40 to 44
years - Counts - Total">  2,730</td>
Number of tables found: 1
```

```
<td class="text-right text-nowrap" headers="rh2 r19 geo2021A0011M1T
geo2021A0011M1Tstat1 geo2021A0011M1Tstat1gen1" title="2021A0011M1T - 40 to 44
years - Counts - Total">  1,840</td>
```

[29]:
```
df
```

[29]:

	postal_code	data	value
0	M3A	2021A0011M3A - 40 to 44 years - Counts - Total	2340.0
1	M4A	2021A0011M4A - 40 to 44 years - Counts - Total	1020.0
2	M5A	2021A0011M5A - 40 to 44 years - Counts - Total	3570.0
3	M6A	2021A0011M6A - 40 to 44 years - Counts - Total	1585.0
4	M7A	2021A0011M7A - 40 to 44 years - Counts - Total	NaN
..	…	…	…
69	M1X	2021A0011M1X - 40 to 44 years - Counts - Total	890.0
70	M4X	2021A0011M4X - 40 to 44 years - Counts - Total	1505.0
71	M8X	2021A0011M8X - 40 to 44 years - Counts - Total	605.0
72	M8Y	2021A0011M8Y - 40 to 44 years - Counts - Total	1745.0
73	M8Z	2021A0011M8Z - 40 to 44 years - Counts - Total	1590.0

[74 rows x 3 columns]

[30]:
```
# Now you can export this to a CSV file for further analysis or visulization
df.to_csv('40to44.csv')
```

[ ]:
```

```