Simple BeautifulSoup examplenew

November 22, 2024

1 Simple Webscraping Example with Beautiful Soup

```
[1]: # Import all neccessary libraries
     from bs4 import BeautifulSoup
     import urllib.request
     import pandas as pd
[2]: # 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
[2]: <http.client.HTTPResponse at 0x7f1526f27160>
[3]: # Create the BeautifulSoup object
    html_to_parse = BeautifulSoup(html, "html.parser")
[4]: # 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
[5]: # Create list of all the > tags in the table that has the title_
      →"2021A0011M1C - Population, 2021 - Counts - Total"
```

```
td = tables[0].find(attrs={"title":"2021A0011M1C - Population, 2021 - Counts -⊔

Government of the stable of the
```

[6]: 35,642

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

[7]: 35642.0

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

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

```
[9]: # A list of postal code from the previous part
    postal = ['M3A', 'M4A', 'M5A', 'M6A', 'M7A', 'M5E', 'M4E', 'M6E', 'M5G',
     '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',
```

```
'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',]
[10]: # Creating Empty DataFrame and Storing it in variable df
      df = pd.DataFrame(columns = ['postal_code', 'data', 'value'])
[15]: # Loop through each postal code
      for i in postal:
```

'M6N', 'M9N',

```
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} - Population, 2021 - Counts - Total")
    td = tables[0].find(attrs={"title":title})
    print(td)
    df.loc[len(df.index)] = [i, title, float(td.text.replace(",", ""))]
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M3A
geo2021A0011M3Astat1 geo2021A0011M3Astat1gen1" title="2021A0011M3A - Population,
2021 - Counts - Total"> 34.361
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M4A
geo2021A0011M4Astat1 geo2021A0011M4Astat1gen1" title="2021A0011M4A - Population,
2021 - Counts - Total"> 14,589
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M5A
geo2021A0011M5Astat1 geo2021A0011M5Astat1gen1" title="2021A0011M5A - Population,
2021 - Counts - Total"> 48,978
Number of tables found: 1
```

```
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M6A
geo2021A0011M6Astat1 geo2021A0011M6Astat1gen1" title="2021A0011M6A - Population,
2021 - Counts - Total"> 22,380
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M7A
geo2021A0011M7Astat1 geo2021A0011M7Astat1gen1" title="2021A0011M7A - Population,
2021 - Counts - Total">
                        5
Number of tables found: 1
geo2021A0011M5Estat1 geo2021A0011M5Estat1gen1" title="2021A0011M5E - Population,
2021 - Counts - Total"> 11,779
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M4E
geo2021A0011M4Estat1 geo2021A0011M4Estat1gen1" title="2021A0011M4E - Population,
2021 - Counts - Total"> 25,473
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M6E
geo2021A0011M6Estat1 geo2021A0011M6Estat1gen1" title="2021A0011M6E - Population,
2021 - Counts - Total"> 37,684
Number of tables found: 1
geo2021A0011M5Gstat1 geo2021A0011M5Gstat1gen1" title="2021A0011M5G - Population,
2021 - Counts - Total"> 9.751
Number of tables found: 1
geo2021A0011M6Gstat1 geo2021A0011M6Gstat1gen1" title="2021A0011M6G - Population,
2021 - Counts - Total"> 30,600
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M2H
geo2021A0011M2Hstat1 geo2021A0011M2Hstat1gen1" title="2021A0011M2H - Population,
2021 - Counts - Total"> 23,396
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M3H
geo2021A0011M3Hstat1 geo2021A0011M3Hstat1gen1" title="2021A0011M3H - Population,
2021 - Counts - Total"> 38,416
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M4H
geo2021A0011M4Hstat1 geo2021A0011M4Hstat1gen1" title="2021A0011M4H - Population,
2021 - Counts - Total"> 18,698
Number of tables found: 1
geo2021A0011M5Hstat1 geo2021A0011M5Hstat1gen1" title="2021A0011M5H - Population,
2021 - Counts - Total"> 3,248
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M6H
geo2021A0011M6Hstat1 geo2021A0011M6Hstat1gen1" title="2021A0011M6H - Population,
2021 - Counts - Total"> 43,730
Number of tables found: 1
```

```
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M1J
geo2021A0011M1Jstat1 geo2021A0011M1Jstat1gen1" title="2021A0011M1J - Population,
2021 - Counts - Total"> 37,002
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M2J
geo2021A0011M2Jstat1 geo2021A0011M2Jstat1gen1" title="2021A0011M2J - Population,
2021 - Counts - Total"> 61,761
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M3J
geo2021A0011M3Jstat1 geo2021A0011M3Jstat1gen1" title="2021A0011M3J - Population,
2021 - Counts - Total"> 26,600
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M4J
geo2021A0011M4Jstat1 geo2021A0011M4Jstat1gen1" title="2021A0011M4J - Population,
2021 - Counts - Total"> 35,128
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M5J
geo2021A0011M5Jstat1 geo2021A0011M5Jstat1gen1" title="2021A0011M5J - Population,
2021 - Counts - Total"> 16,879
Number of tables found: 0
No tables found on this page.
Number of tables found: 1
geo2021A0011M6Kstat1 geo2021A0011M6Kstat1gen1" title="2021A0011M6K - Population,
2021 - Counts - Total"> 42,110
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M1L
geo2021A0011M1Lstat1 geo2021A0011M1Lstat1gen1" title="2021A0011M1L - Population,
2021 - Counts - Total"> 35,833
Number of tables found: 1
geo2021A0011M2Lstat1 geo2021A0011M2Lstat1gen1" title="2021A0011M2L - Population,
2021 - Counts - Total"> 11,274
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M3L
geo2021A0011M3Lstat1 geo2021A0011M3Lstat1gen1" title="2021A0011M3L - Population,
2021 - Counts - Total"> 19,263
Number of tables found: 1
geo2021A0011M4Lstat1 geo2021A0011M4Lstat1gen1" title="2021A0011M4L - Population,
2021 - Counts - Total"> 32,218
Number of tables found: 0
No tables found on this page.
Number of tables found: 1
geo2021A0011M6Lstat1 geo2021A0011M6Lstat1gen1" title="2021A0011M6L - Population,
2021 - Counts - Total"> 20,156
Number of tables found: 1
```

```
geo2021A0011M9Lstat1 geo2021A0011M9Lstat1gen1" title="2021A0011M9L - Population,
2021 - Counts - Total"> 11,737
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M1M
geo2021A0011M1Mstat1 geo2021A0011M1Mstat1gen1" title="2021A0011M1M - Population,
2021 - Counts - Total"> 23,258
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M2M
geo2021A0011M2Mstat1 geo2021A0011M2Mstat1gen1" title="2021A0011M2M - Population,
2021 - Counts - Total"> 31,915
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M3M
geo2021A0011M3Mstat1 geo2021A0011M3Mstat1gen1" title="2021A0011M3M - Population,
2021 - Counts - Total"> 24,906
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M4M
geo2021A0011M4Mstat1 geo2021A0011M4Mstat1gen1" title="2021A0011M4M - Population,
2021 - Counts - Total"> 25,052
Number of tables found: 1
geo2021A0011M5Mstat1 geo2021A0011M5Mstat1gen1" title="2021A0011M5M - Population,
2021 - Counts - Total"> 25,495
Number of tables found: 1
geo2021A0011M6Mstat1 geo2021A0011M6Mstat1gen1" title="2021A0011M6M - Population,
2021 - Counts - Total"> 42,148
Number of tables found: 1
geo2021A0011M9Mstat1 geo2021A0011M9Mstat1gen1" title="2021A0011M9M - Population,
2021 - Counts - Total"> 24,245
Number of tables found: 1
geo2021A0011M1Nstat1 geo2021A0011M1Nstat1gen1" title="2021A0011M1N - Population,
2021 - Counts - Total"> 22,976
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M2N
geo2021A0011M2Nstat1 geo2021A0011M2Nstat1gen1" title="2021A0011M2N - Population,
2021 - Counts - Total"> 75,100
Number of tables found: 1
geo2021A0011M3Nstat1 geo2021A0011M3Nstat1gen1" title="2021A0011M3N - Population,
2021 - Counts - Total"> 40,846
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M4N
geo2021A0011M4Nstat1 geo2021A0011M4Nstat1gen1" title="2021A0011M4N - Population,
2021 - Counts - Total"> 16,058
Number of tables found: 1
```

```
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M5N
geo2021A0011M5Nstat1 geo2021A0011M5Nstat1gen1" title="2021A0011M5N - Population,
2021 - Counts - Total"> 16,154
Number of tables found: 1
geo2021A0011M6Nstat1 geo2021A0011M6Nstat1gen1" title="2021A0011M6N - Population,
2021 - Counts - Total"> 40,252
Number of tables found: 1
geo2021A0011M9Nstat1 geo2021A0011M9Nstat1gen1" title="2021A0011M9N - Population,
2021 - Counts - Total"> 26,059
Number of tables found: 1
geo2021A0011M1Pstat1 geo2021A0011M1Pstat1gen1" title="2021A0011M1P - Population,
2021 - Counts - Total"> 45,170
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M2P
geo2021A0011M2Pstat1 geo2021A0011M2Pstat1gen1" title="2021A0011M2P - Population,
2021 - Counts - Total"> 7,546
Number of tables found: 1
geo2021A0011M4Pstat1 geo2021A0011M4Pstat1gen1" title="2021A0011M4P - Population,
2021 - Counts - Total"> 25,057
Number of tables found: 1
geo2021A0011M5Pstat1 geo2021A0011M5Pstat1gen1" title="2021A0011M5P - Population,
2021 - Counts - Total"> 19,791
Number of tables found: 1
geo2021A0011M6Pstat1 geo2021A0011M6Pstat1gen1" title="2021A0011M6P - Population,
2021 - Counts - Total"> 39,870
Number of tables found: 1
geo2021A0011M9Pstat1 geo2021A0011M9Pstat1gen1" title="2021A0011M9P - Population,
2021 - Counts - Total"> 20,645
Number of tables found: 1
geo2021A0011M1Rstat1 geo2021A0011M1Rstat1gen1" title="2021A0011M1R - Population,
2021 - Counts - Total"> 30,467
Number of tables found: 1
geo2021A0011M2Rstat1 geo2021A0011M2Rstat1gen1" title="2021A0011M2R - Population,
2021 - Counts - Total"> 40,581
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M4R
geo2021A0011M4Rstat1 geo2021A0011M4Rstat1gen1" title="2021A0011M4R - Population,
2021 - Counts - Total"> 11,909
Number of tables found: 1
```

```
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M5R
geo2021A0011M5Rstat1 geo2021A0011M5Rstat1gen1" title="2021A0011M5R - Population,
2021 - Counts - Total"> 26,197
Number of tables found: 1
geo2021A0011M6Rstat1 geo2021A0011M6Rstat1gen1" title="2021A0011M6R - Population,
2021 - Counts - Total"> 19,349
Number of tables found: 0
No tables found on this page.
Number of tables found: 1
geo2021A0011M9Rstat1 geo2021A0011M9Rstat1gen1" title="2021A0011M9R - Population,
2021 - Counts - Total"> 33,783
Number of tables found: 1
geo2021A0011M1Sstat1 geo2021A0011M1Sstat1gen1" title="2021A0011M1S - Population,
2021 - Counts - Total"> 37,663
Number of tables found: 1
geo2021A0011M4Sstat1 geo2021A0011M4Sstat1gen1" title="2021A0011M4S - Population,
2021 - Counts - Total"> 30,754
Number of tables found: 1
geo2021A0011M5Sstat1 geo2021A0011M5Sstat1gen1" title="2021A0011M5S - Population,
2021 - Counts - Total"> 17,074
Number of tables found: 1
geo2021A0011M6Sstat1 geo2021A0011M6Sstat1gen1" title="2021A0011M6S - Population,
2021 - Counts - Total"> 34,029
Number of tables found: 1
geo2021A0011M1Tstat1 geo2021A0011M1Tstat1gen1" title="2021A0011M1T - Population,
2021 - Counts - Total"> 34,607
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M4T
geo2021A0011M4Tstat1 geo2021A0011M4Tstat1gen1" title="2021A0011M4T - Population,
2021 - Counts - Total"> 10,332
Number of tables found: 1
geo2021A0011M5Tstat1 geo2021A0011M5Tstat1gen1" title="2021A0011M5T - Population,
2021 - Counts - Total"> 17,903
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M1V
geo2021A0011M1Vstat1 geo2021A0011M1Vstat1gen1" title="2021A0011M1V - Population,
2021 - Counts - Total"> 50,825
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M4V
geo2021A0011M4Vstat1 geo2021A0011M4Vstat1gen1" title="2021A0011M4V - Population,
```

```
2021 - Counts - Total"> 19,273
Number of tables found: 1
geo2021A0011M5Vstat1 geo2021A0011M5Vstat1gen1" title="2021A0011M5V - Population,
2021 - Counts - Total"> 59,912
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M8V
geo2021A0011M8Vstat1 geo2021A0011M8Vstat1gen1" title="2021A0011M8V - Population,
2021 - Counts - Total"> 44,144
Number of tables found: 1
geo2021A0011M9Vstat1 geo2021A0011M9Vstat1gen1" title="2021A0011M9V - Population,
2021 - Counts - Total"> 53,878
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M1W
geo2021A0011M1Wstat1 geo2021A0011M1Wstat1gen1" title="2021A0011M1W - Population,
2021 - Counts - Total"> 47,070
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M4W
geo2021A0011M4Wstat1 geo2021A0011M4Wstat1gen1" title="2021A0011M4W - Population,
2021 - Counts - Total"> 15,296
Number of tables found: 0
No tables found on this page.
Number of tables found: 1
geo2021A0011M8Wstat1 geo2021A0011M8Wstat1gen1" title="2021A0011M8W - Population,
2021 - Counts - Total"> 22,381
Number of tables found: 1
geo2021A0011M9Wstat1 geo2021A0011M9Wstat1gen1" title="2021A0011M9W - Population,
2021 - Counts - Total"> 40,977
Number of tables found: 1
geo2021A0011M1Xstat1 geo2021A0011M1Xstat1gen1" title="2021A0011M1X - Population,
2021 - Counts - Total"> 14,810
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M4X
geo2021A0011M4Xstat1 geo2021A0011M4Xstat1gen1" title="2021A0011M4X - Population,
2021 - Counts - Total"> 19,896
Number of tables found: 0
No tables found on this page.
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M8X
geo2021A0011M8Xstat1 geo2021A0011M8Xstat1gen1" title="2021A0011M8X - Population,
2021 - Counts - Total"> 10,624
Number of tables found: 0
No tables found on this page.
Number of tables found: 1
```

```
geo2021A0011M8Ystat1 geo2021A0011M8Ystat1gen1" title="2021A0011M8Y - Population,
     2021 - Counts - Total"> 21,986
     Number of tables found: 1
     <td class="text-right text-nowrap" headers="rh1 r1 geo2021A0011M8Z
     geo2021A0011M8Zstat1 geo2021A0011M8Zstat1gen1" title="2021A0011M8Z - Population,
     2021 - Counts - Total"> 19,312
[16]: df
[16]:
        postal code
                                                                     value
               M3A 2021A0011M3A - Population, 2021 - Counts - Total 34361.0
               M4A 2021A0011M4A - Population, 2021 - Counts - Total
                                                                   14589.0
     1
               M5A 2021A0011M5A - Population, 2021 - Counts - Total 48978.0
     2
               M6A 2021A0011M6A - Population, 2021 - Counts - Total
     3
                                                                   22380.0
               M7A 2021A0011M7A - Population, 2021 - Counts - Total
     4
                                                                      5.0
               M1X 2021A0011M1X - Population, 2021 - Counts - Total 14810.0
     89
               M4X 2021A0011M4X - Population, 2021 - Counts - Total
     90
                                                                   19896.0
               M8X 2021A0011M8X - Population, 2021 - Counts - Total
     91
                                                                   10624.0
               M8Y 2021A0011M8Y - Population, 2021 - Counts - Total 21986.0
     92
     93
               M8Z 2021A0011M8Z - Population, 2021 - Counts - Total 19312.0
     [94 rows x 3 columns]
[17]: # Now you can export this to a CSV file for further analysis or visulization
     df.to_csv('postal.csv')
[]:
```