house -Copy2

November 24, 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]: <a href="http.client.HTTPResponse">http.client.HTTPResponse</a> at 0x7f24dbe9f790>
[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_\mathbb{\text{\text{l}}}
      →"2021A0011M1C - Population, 2021 - Counts - Total"
```

```
td = tables[0].find(attrs={"title":"2021A0011M1C - 25 to 29 years - Counts -□ →Total"})

[6]: td

[6]:  2,385
```

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

[7]: 2385.0

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',
```

```
'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',
              'W8V',
              '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'])
[11]: # 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} - 25 to 29 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 r16 geo2021A0011M3A
geo2021A0011M3Astat1 geo2021A0011M3Astat1gen1" title="2021A0011M3A - 25 to 29
vears - Counts - Total"> 2,525
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M4A
geo2021A0011M4Astat1 geo2021A0011M4Astat1gen1" title="2021A0011M4A - 25 to 29
years - Counts - Total">
                       920
Number of tables found: 1
geo2021A0011M5Astat1 geo2021A0011M5Astat1gen1" title="2021A0011M5A - 25 to 29
years - Counts - Total"> 6,310
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M6A
geo2021A0011M6Astat1 geo2021A0011M6Astat1gen1" title="2021A0011M6A - 25 to 29
years - Counts - Total"> 1,765
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M7A
geo2021A0011M7Astat1 geo2021A0011M7Astat1gen1" title="2021A0011M7A - 25 to 29
years - Counts - Total"><abbr title="suppressed to meet the confidentiality
requirements of the Statistics Act">x</abbr>
Data suppressed or invalid for M7A, setting value to NaN.
Number of tables found: 1
geo2021A0011M5Estat1 geo2021A0011M5Estat1gen1" title="2021A0011M5E - 25 to 29
years - Counts - Total"> 2,375
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M4E
geo2021A0011M4Estat1 geo2021A0011M4Estat1gen1" title="2021A0011M4E - 25 to 29
years - Counts - Total"> 1,105
Number of tables found: 1
geo2021A0011M6Estat1 geo2021A0011M6Estat1gen1" title="2021A0011M6E - 25 to 29
years - Counts - Total"> 2,840
Number of tables found: 1
geo2021A0011M5Gstat1 geo2021A0011M5Gstat1gen1" title="2021A0011M5G - 25 to 29
years - Counts - Total"> 2,130
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M6G
geo2021A0011M6Gstat1 geo2021A0011M6Gstat1gen1" title="2021A0011M6G - 25 to 29
years - Counts - Total"> 2,995
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M2H
geo2021A0011M2Hstat1 geo2021A0011M2Hstat1gen1" title="2021A0011M2H - 25 to 29
```

```
years - Counts - Total"> 1,595
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M3H
geo2021A0011M3Hstat1 geo2021A0011M3Hstat1gen1" title="2021A0011M3H - 25 to 29
years - Counts - Total"> 2,680
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M4H
geo2021A0011M4Hstat1 geo2021A0011M4Hstat1gen1" title="2021A0011M4H - 25 to 29
years - Counts - Total"> 1,245
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M5H
geo2021A0011M5Hstat1 geo2021A0011M5Hstat1gen1" title="2021A0011M5H - 25 to 29
years - Counts - Total">
                        785
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M6H
geo2021A0011M6Hstat1 geo2021A0011M6Hstat1gen1" title="2021A0011M6H - 25 to 29
years - Counts - Total"> 4,260
Number of tables found: 1
geo2021A0011M1Jstat1 geo2021A0011M1Jstat1gen1" title="2021A0011M1J - 25 to 29
years - Counts - Total"> 2,855
Number of tables found: 1
geo2021A0011M2Jstat1 geo2021A0011M2Jstat1gen1" title="2021A0011M2J - 25 to 29
years - Counts - Total"> 6,220
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M3J
geo2021A0011M3Jstat1 geo2021A0011M3Jstat1gen1" title="2021A0011M3J - 25 to 29
years - Counts - Total"> 2,760
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M4J
geo2021A0011M4Jstat1 geo2021A0011M4Jstat1gen1" title="2021A0011M4J - 25 to 29
years - Counts - Total"> 1,910
Number of tables found: 1
geo2021A0011M5Jstat1 geo2021A0011M5Jstat1gen1" title="2021A0011M5J - 25 to 29
years - Counts - Total"> 3,365
Number of tables found: 0
No tables found on this page.
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M6K
geo2021A0011M6Kstat1 geo2021A0011M6Kstat1gen1" title="2021A0011M6K - 25 to 29
years - Counts - Total"> 6,220
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M1L
geo2021A0011M1Lstat1 geo2021A0011M1Lstat1gen1" title="2021A0011M1L - 25 to 29
years - Counts - Total"> 2,435
Number of tables found: 1
```

```
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M2L
geo2021A0011M2Lstat1 geo2021A0011M2Lstat1gen1" title="2021A0011M2L - 25 to 29
years - Counts - Total">
                         655
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M3L
geo2021A0011M3Lstat1 geo2021A0011M3Lstat1gen1" title="2021A0011M3L - 25 to 29
years - Counts - Total"> 1,285
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M4L
geo2021A0011M4Lstat1 geo2021A0011M4Lstat1gen1" title="2021A0011M4L - 25 to 29
years - Counts - Total"> 1,665
Number of tables found: 0
No tables found on this page.
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M6L
geo2021A0011M6Lstat1 geo2021A0011M6Lstat1gen1" title="2021A0011M6L - 25 to 29
years - Counts - Total"> 1,365
Number of tables found: 1
geo2021A0011M9Lstat1 geo2021A0011M9Lstat1gen1" title="2021A0011M9L - 25 to 29
years - Counts - Total">
                         950
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M1M
geo2021A0011M1Mstat1 geo2021A0011M1Mstat1gen1" title="2021A0011M1M - 25 to 29
years - Counts - Total"> 1,450
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M2M
geo2021A0011M2Mstat1 geo2021A0011M2Mstat1gen1" title="2021A0011M2M - 25 to 29
years - Counts - Total"> 2,735
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M3M
\tt geo2021A0011M3Mstat1\ geo2021A0011M3Mstat1gen1"\ title="2021A0011M3M\ -\ 25\ to\ 29
years - Counts - Total"> 1,835
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M4M
geo2021A0011M4Mstat1 geo2021A0011M4Mstat1gen1" title="2021A0011M4M - 25 to 29
years - Counts - Total"> 2,190
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M5M
geo2021A0011M5Mstat1 geo2021A0011M5Mstat1gen1" title="2021A0011M5M - 25 to 29
years - Counts - Total">
                         950
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M6M
geo2021A0011M6Mstat1 geo2021A0011M6Mstat1gen1" title="2021A0011M6M - 25 to 29
years - Counts - Total"> 2,955
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M9M
geo2021A0011M9Mstat1 geo2021A0011M9Mstat1gen1" title="2021A0011M9M - 25 to 29
```

```
years - Counts - Total"> 1,785
Number of tables found: 1
geo2021A0011M1Nstat1 geo2021A0011M1Nstat1gen1" title="2021A0011M1N - 25 to 29
years - Counts - Total"> 1,345
Number of tables found: 1
geo2021A0011M2Nstat1 geo2021A0011M2Nstat1gen1" title="2021A0011M2N - 25 to 29
years - Counts - Total"> 7,955
Number of tables found: 1
geo2021A0011M3Nstat1 geo2021A0011M3Nstat1gen1" title="2021A0011M3N - 25 to 29
years - Counts - Total"> 2,910
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M4N
geo2021A0011M4Nstat1 geo2021A0011M4Nstat1gen1" title="2021A0011M4N - 25 to 29
years - Counts - Total">
                   715
Number of tables found: 1
geo2021A0011M5Nstat1 geo2021A0011M5Nstat1gen1" title="2021A0011M5N - 25 to 29
years - Counts - Total">
                   870
Number of tables found: 1
geo2021A0011M6Nstat1 geo2021A0011M6Nstat1gen1" title="2021A0011M6N - 25 to 29
years - Counts - Total"> 2,900
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M9N
geo2021A0011M9Nstat1 geo2021A0011M9Nstat1gen1" title="2021A0011M9N - 25 to 29
years - Counts - Total"> 1,885
Number of tables found: 1
geo2021A0011M1Pstat1 geo2021A0011M1Pstat1gen1" title="2021A0011M1P - 25 to 29
years - Counts - Total"> 3,735
Number of tables found: 1
geo2021A0011M2Pstat1 geo2021A0011M2Pstat1gen1" title="2021A0011M2P - 25 to 29
years - Counts - Total">
                   400
Number of tables found: 1
geo2021A0011M4Pstat1 geo2021A0011M4Pstat1gen1" title="2021A0011M4P - 25 to 29
years - Counts - Total"> 3,110
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M5P
geo2021A0011M5Pstat1 geo2021A0011M5Pstat1gen1" title="2021A0011M5P - 25 to 29
years - Counts - Total"> 1,585
Number of tables found: 1
geo2021A0011M6Pstat1 geo2021A0011M6Pstat1gen1" title="2021A0011M6P - 25 to 29
```

```
years - Counts - Total"> 3,255
Number of tables found: 1
geo2021A0011M9Pstat1 geo2021A0011M9Pstat1gen1" title="2021A0011M9P - 25 to 29
years - Counts - Total"> 1,130
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M1R
geo2021A0011M1Rstat1 geo2021A0011M1Rstat1gen1" title="2021A0011M1R - 25 to 29
years - Counts - Total"> 2,370
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M2R
geo2021A0011M2Rstat1 geo2021A0011M2Rstat1gen1" title="2021A0011M2R - 25 to 29
years - Counts - Total"> 2,800
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M4R
geo2021A0011M4Rstat1 geo2021A0011M4Rstat1gen1" title="2021A0011M4R - 25 to 29
years - Counts - Total">
                     825
Number of tables found: 1
geo2021A0011M5Rstat1 geo2021A0011M5Rstat1gen1" title="2021A0011M5R - 25 to 29
years - Counts - Total"> 2,940
Number of tables found: 1
geo2021A0011M6Rstat1 geo2021A0011M6Rstat1gen1" title="2021A0011M6R - 25 to 29
years - Counts - Total"> 1,390
Number of tables found: 0
No tables found on this page.
Number of tables found: 1
geo2021A0011M9Rstat1 geo2021A0011M9Rstat1gen1" title="2021A0011M9R - 25 to 29
years - Counts - Total"> 2,285
Number of tables found: 1
geo2021A0011M1Sstat1 geo2021A0011M1Sstat1gen1" title="2021A0011M1S - 25 to 29
years - Counts - Total"> 3,250
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M4S
geo2021A0011M4Sstat1 geo2021A0011M4Sstat1gen1" title="2021A0011M4S - 25 to 29
years - Counts - Total"> 3,165
Number of tables found: 1
geo2021A0011M5Sstat1 geo2021A0011M5Sstat1gen1" title="2021A0011M5S - 25 to 29
years - Counts - Total"> 2,810
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M6S
geo2021A0011M6Sstat1 geo2021A0011M6Sstat1gen1" title="2021A0011M6S - 25 to 29
years - Counts - Total"> 1,675
Number of tables found: 1
```

```
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M1T
geo2021A0011M1Tstat1 geo2021A0011M1Tstat1gen1" title="2021A0011M1T - 25 to 29
years - Counts - Total"> 2,515
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M4T
geo2021A0011M4Tstat1 geo2021A0011M4Tstat1gen1" title="2021A0011M4T - 25 to 29
years - Counts - Total">
                       610
Number of tables found: 1
geo2021A0011M5Tstat1 geo2021A0011M5Tstat1gen1" title="2021A0011M5T - 25 to 29
years - Counts - Total"> 2,710
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M1V
geo2021A0011M1Vstat1 geo2021A0011M1Vstat1gen1" title="2021A0011M1V - 25 to 29
years - Counts - Total"> 3,390
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M4V
geo2021A0011M4Vstat1 geo2021A0011M4Vstat1gen1" title="2021A0011M4V - 25 to 29
years - Counts - Total"> 1,580
Number of tables found: 1
geo2021A0011M5Vstat1 geo2021A0011M5Vstat1gen1" title="2021A0011M5V - 25 to 29
years - Counts - Total"> 14,550
Number of tables found: 1
geo2021A0011M8Vstat1 geo2021A0011M8Vstat1gen1" title="2021A0011M8V - 25 to 29
years - Counts - Total"> 4,250
Number of tables found: 1
geo2021A0011M9Vstat1 geo2021A0011M9Vstat1gen1" title="2021A0011M9V - 25 to 29
years - Counts - Total"> 4,730
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M1W
geo2021A0011M1Wstat1 geo2021A0011M1Wstat1gen1" title="2021A0011M1W - 25 to 29
years - Counts - Total"> 3,140
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M4W
geo2021A0011M4Wstat1 geo2021A0011M4Wstat1gen1" title="2021A0011M4W - 25 to 29
years - Counts - Total"> 1,385
Number of tables found: 0
No tables found on this page.
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M8W
geo2021A0011M8Wstat1 geo2021A0011M8Wstat1gen1" title="2021A0011M8W - 25 to 29
years - Counts - Total"> 1,490
Number of tables found: 1
<td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M9W
geo2021A0011M9Wstat1 geo2021A0011M9Wstat1gen1" title="2021A0011M9W - 25 to 29
```

```
years - Counts - Total"> 3,760
    Number of tables found: 1
    geo2021A0011M1Xstat1 geo2021A0011M1Xstat1gen1" title="2021A0011M1X - 25 to 29
    years - Counts - Total"> 1,200
    Number of tables found: 1
    <td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M4X
    geo2021A0011M4Xstat1 geo2021A0011M4Xstat1gen1" title="2021A0011M4X - 25 to 29
    years - Counts - Total"> 1,795
    Number of tables found: 0
    No tables found on this page.
    Number of tables found: 1
    <td class="text-right text-nowrap" headers="rh2 r16 geo2021A0011M8X
    geo2021A0011M8Xstat1 geo2021A0011M8Xstat1gen1" title="2021A0011M8X - 25 to 29
    years - Counts - Total">
                            425
    Number of tables found: 0
    No tables found on this page.
    Number of tables found: 1
    geo2021A0011M8Ystat1 geo2021A0011M8Ystat1gen1" title="2021A0011M8Y - 25 to 29
    years - Counts - Total"> 1,545
    Number of tables found: 1
    geo2021A0011M8Zstat1 geo2021A0011M8Zstat1gen1" title="2021A0011M8Z - 25 to 29
    years - Counts - Total"> 1,455
[12]: df
[12]:
       postal_code
                                                        data
                                                              value
     0
                   2021A0011M3A - 25 to 29 years - Counts - Total
                                                             2525.0
              MSA
              M4A 2021A0011M4A - 25 to 29 years - Counts - Total
     1
                                                              920.0
     2
              M5A 2021A0011M5A - 25 to 29 years - Counts - Total 6310.0
     3
              M6A 2021A0011M6A - 25 to 29 years - Counts - Total
                                                             1765.0
              M7A 2021A0011M7A - 25 to 29 years - Counts - Total
                                                                NaN
              M1X 2021A0011M1X - 25 to 29 years - Counts - Total 1200.0
     69
     70
              M4X 2021A0011M4X - 25 to 29 years - Counts - Total 1795.0
     71
              M8X 2021A0011M8X - 25 to 29 years - Counts - Total
                                                              425.0
              M8Y 2021A0011M8Y - 25 to 29 years - Counts - Total 1545.0
     72
              M8Z 2021A0011M8Z - 25 to 29 years - Counts - Total 1455.0
     73
     [74 rows x 3 columns]
[14]: # Now you can export this to a CSV file for further analysis or visulization
     df.to_csv('25to29.csv')
```

[]: