Final Assignment_Webscraping

March 5, 2023

Extracting Stock Data Using a Web Scraping

Not all stock data is available via API in this assignment; you will use web-scraping to obtain financial data. You will be quizzed on your results.
Using beautiful soup we will extract historical share data from a web-page.

Table of Contents

```
<ul>
    <li>Downloading the Webpage Using Requests Library</li>
    <li>Parsing Webpage HTML Using BeautifulSoup</li>
    <li>Extracting Data and Building DataFrame</li>
</ul>
```

Estimated Time Needed: 30 min

[1]:

```
#!pip install pandas==1.3.3
#!pip install requests==2.26.0
!mamba install bs4==4.10.0 -y
!mamba install html5lib==1.1 -y
!pip install lxml==4.6.4
#!pip install plotly==5.3.1
```

```
          __    __    __    __
         /  \  /  \  /  \  /  \
        /    \/    \/    \/    \
/  / /  / /  / /  /
      /  / \   / \   / \   / \  \____
     /  /   \_/   \_/   \_/   \    o \__,
    / _/                       \_____/  `
    |/




mamba (0.15.3) supported by @QuantStack
```

```
        GitHub:  https://github.com/mamba-org/mamba
        Twitter: https://twitter.com/QuantStack



Looking for: ['bs4==4.10.0']

pkgs/main/linux-64       [<=>                 ] (00m:00s)
pkgs/main/linux-64       [=>                ] (00m:00s) 436 KB / ?? (1.35 MB/s)
pkgs/main/linux-64       [=>                ] (00m:00s) 436 KB / ?? (1.35 MB/s)
pkgs/main/noarch         [<=>                 ] (00m:00s)
pkgs/main/linux-64       [=>                ] (00m:00s) 436 KB / ?? (1.35 MB/s)
pkgs/main/noarch         [=>                ] (00m:00s) 352 KB / ?? (1.08 MB/s)
pkgs/main/linux-64       [=>                ] (00m:00s) 436 KB / ?? (1.35 MB/s)
pkgs/main/noarch         [=>                ] (00m:00s) 352 KB / ?? (1.08 MB/s)
pkgs/r/linux-64          [<=>                 ] (00m:00s)
pkgs/main/linux-64       [=>                ] (00m:00s) 436 KB / ?? (1.35 MB/s)
pkgs/main/noarch         [=>                ] (00m:00s) 352 KB / ?? (1.08 MB/s)
pkgs/r/linux-64          [=>                ] (00m:00s) 640 KB / ?? (1.96 MB/s)
pkgs/main/linux-64       [=>                ] (00m:00s) 436 KB / ?? (1.35 MB/s)
pkgs/main/noarch         [=>                ] (00m:00s) 352 KB / ?? (1.08 MB/s)
pkgs/r/linux-64          [=>                ] (00m:00s) 640 KB / ?? (1.96 MB/s)
pkgs/r/noarch            [<=>                 ] (00m:00s)
pkgs/main/linux-64       [=>                ] (00m:00s) 436 KB / ?? (1.35 MB/s)
pkgs/main/noarch         [=>                ] (00m:00s) 352 KB / ?? (1.08 MB/s)
pkgs/r/linux-64          [=>                ] (00m:00s) 640 KB / ?? (1.96 MB/s)
pkgs/r/noarch            [=>                ] (00m:00s) 696 KB / ?? (2.12 MB/s)
pkgs/main/linux-64       [=>                ] (00m:00s) 436 KB / ?? (1.35 MB/s)
pkgs/main/noarch         [=>                ] (00m:00s) 352 KB / ?? (1.08 MB/s)
pkgs/r/linux-64          [=>                ] (00m:00s) 640 KB / ?? (1.96 MB/s)
pkgs/r/noarch            [<=>                 ] (00m:00s) Finalizing…
pkgs/main/linux-64       [=>                ] (00m:00s) 436 KB / ?? (1.35 MB/s)
pkgs/main/noarch         [=>                ] (00m:00s) 352 KB / ?? (1.08 MB/s)
pkgs/r/linux-64          [=>                ] (00m:00s) 640 KB / ?? (1.96 MB/s)
pkgs/r/noarch            [<=>                 ] (00m:00s) Done
pkgs/r/noarch            [====================] (00m:00s) Done
pkgs/main/linux-64       [=>                ] (00m:00s) 436 KB / ?? (1.35 MB/s)
pkgs/main/noarch         [=>                ] (00m:00s) 352 KB / ?? (1.08 MB/s)
pkgs/r/linux-64          [=>                ] (00m:00s) 640 KB / ?? (1.96 MB/s)
pkgs/main/linux-64       [<=>               ] (00m:00s) 436 KB / ?? (1.35 MB/s)
pkgs/main/noarch         [=>                ] (00m:00s) 352 KB / ?? (1.08 MB/s)
pkgs/r/linux-64          [=>                ] (00m:00s) 640 KB / ?? (1.96 MB/s)
pkgs/main/linux-64       [ <=>                ] (00m:00s) 1 MB / ?? (2.19 MB/s)
pkgs/main/noarch         [=>                ] (00m:00s) 352 KB / ?? (1.08 MB/s)
pkgs/r/linux-64          [=>                ] (00m:00s) 640 KB / ?? (1.96 MB/s)
pkgs/main/linux-64       [ <=>                ] (00m:00s) 1 MB / ?? (2.19 MB/s)
```

```
pkgs/main/noarch         [=>                ] (00m:00s) 352 KB / ?? (1.08 MB/s)
pkgs/r/linux-64          [<=>               ] (00m:00s) 640 KB / ?? (1.96 MB/s)
pkgs/main/linux-64       [ <=>                ] (00m:00s) 1 MB / ?? (2.19 MB/s)
pkgs/main/noarch         [=>                ] (00m:00s) 352 KB / ?? (1.08 MB/s)
pkgs/r/linux-64          [ <=>                ] (00m:00s) 1 MB / ?? (2.60 MB/s)
pkgs/main/linux-64       [ <=>                ] (00m:00s) 1 MB / ?? (2.19 MB/s)
pkgs/main/noarch         [<=>               ] (00m:00s) 352 KB / ?? (1.08 MB/s)
pkgs/r/linux-64          [ <=>                ] (00m:00s) 1 MB / ?? (2.60 MB/s)
pkgs/main/linux-64       [ <=>                ] (00m:00s) 1 MB / ?? (2.19 MB/s)
pkgs/main/noarch         [<=>               ] (00m:00s) 704 KB / ?? (1.44 MB/s)
pkgs/r/linux-64          [ <=>                ] (00m:00s) 1 MB / ?? (2.60 MB/s)
pkgs/main/linux-64       [ <=>                ] (00m:00s) 1 MB / ?? (2.19 MB/s)
pkgs/main/noarch         [ <=>                ] (00m:00s) Finalizing…
pkgs/r/linux-64          [ <=>                ] (00m:00s) 1 MB / ?? (2.60 MB/s)
pkgs/main/linux-64       [ <=>                ] (00m:00s) 1 MB / ?? (2.19 MB/s)
pkgs/main/noarch         [ <=>                ] (00m:00s) Done
pkgs/r/linux-64          [ <=>                ] (00m:00s) 1 MB / ?? (2.60 MB/s)
pkgs/main/noarch         [====================] (00m:00s) Done
pkgs/main/linux-64       [ <=>                ] (00m:00s) 1 MB / ?? (2.19 MB/s)
pkgs/r/linux-64          [ <=>                ] (00m:00s) 1 MB / ?? (2.60 MB/s)
pkgs/main/linux-64       [ <=>                ] (00m:00s) 1 MB / ?? (2.19 MB/s)
pkgs/r/linux-64          [ <=>                ] (00m:00s) Finalizing…
pkgs/main/linux-64       [ <=>                ] (00m:00s) 1 MB / ?? (2.19 MB/s)
pkgs/r/linux-64          [ <=>                ] (00m:00s) Done
pkgs/r/linux-64          [====================] (00m:00s) Done
pkgs/main/linux-64       [ <=>                ] (00m:00s) 1 MB / ?? (2.19 MB/s)
pkgs/main/linux-64       [  <=>               ] (00m:00s) 1 MB / ?? (2.19 MB/s)
pkgs/main/linux-64       [  <=>               ] (00m:00s) 2 MB / ?? (2.54 MB/s)
pkgs/main/linux-64       [   <=>              ] (00m:00s) 2 MB / ?? (2.54 MB/s)
pkgs/main/linux-64       [   <=>              ] (00m:00s) 3 MB / ?? (3.49 MB/s)
pkgs/main/linux-64       [    <=>             ] (00m:00s) 3 MB / ?? (3.49 MB/s)
pkgs/main/linux-64       [    <=>             ] (00m:00s) 3 MB / ?? (3.64 MB/s)
pkgs/main/linux-64       [     <=>            ] (00m:00s) 3 MB / ?? (3.64 MB/s)
pkgs/main/linux-64       [     <=>            ] (00m:00s) 4 MB / ?? (3.83 MB/s)
pkgs/main/linux-64       [      <=>           ] (00m:00s) 4 MB / ?? (3.83 MB/s)
pkgs/main/linux-64       [      <=>           ] (00m:00s) 5 MB / ?? (3.96 MB/s)
pkgs/main/linux-64       [      <=>           ] (00m:00s) Finalizing…
pkgs/main/linux-64       [      <=>           ] (00m:01s) Done
pkgs/main/linux-64       [====================] (00m:01s) Done

Pinned packages:
  - python 3.7.*


Transaction

  Prefix: /home/jupyterlab/conda/envs/python
  Updating specs:

   - bs4==4.10.0
   - ca-certificates
   - certifi
   - openssl


  Package                 Version  Build           Channel                   Size

  Install:


  + bs4                    4.10.0  hd3eb1b0_0      pkgs/main/noarch
10 KB

  Upgrade:


  - ca-certificates   2022.9.24  ha878542_0      installed
  + ca-certificates  2023.01.10  h06a4308_0      pkgs/main/linux-64
120 KB
  - certifi           2022.9.24  pyhd8ed1ab_0    installed
  + certifi           2022.12.7  py37h06a4308_0  pkgs/main/linux-64
150 KB
  - openssl              1.1.1s  h0b41bf4_1      installed
  + openssl              1.1.1t  h7f8727e_0      pkgs/main/linux-64
4 MB

  Downgrade:


  - beautifulsoup4       4.11.1  pyha770c72_0    installed
  + beautifulsoup4       4.10.0  pyh06a4308_0    pkgs/main/noarch
85 KB

  Summary:

  Install: 1 packages
  Upgrade: 3 packages
  Downgrade: 1 packages

  Total download: 4 MB


Downloading  [>                                        ] (00m:00s)   64.98 KB/s
Extracting   [>                                                      ] (--:--)
```

```
Finished bs4                                  (00m:00s)              10
KB     65 KB/s
Downloading  [>                                        ] (00m:00s)   64.98 KB/s
Extracting   [>                                                      ] (--:--)
Downloading  [>                                        ] (00m:00s)   64.98 KB/s
Extracting   [>                                                      ] (--:--)
Downloading  [>                                        ] (00m:00s)   64.98 KB/s
Extracting   [>                                                      ] (--:--)
Downloading  [=>                                       ] (00m:00s)  806.77 KB/s
Extracting   [>                                                      ] (--:--)
Downloading  [==>                                      ] (00m:00s)    1.29 MB/s
Extracting   [>                                                      ] (--:--)
Downloading  [==>                                      ] (00m:00s)    1.29 MB/s
Extracting   [========>                                ] (00m:00s)        1 / 5
Finished ca-certificates                      (00m:00s)             120
KB    745 KB/s
Downloading  [==>                                      ] (00m:00s)    1.29 MB/s
Extracting   [========>                                ] (00m:00s)        1 / 5
Downloading  [===>                                     ] (00m:00s)    2.17 MB/s
Extracting   [========>                                ] (00m:00s)        1 / 5
Downloading  [===>                                     ] (00m:00s)    2.17 MB/s
Extracting   [========>                                ] (00m:00s)        1 / 5
Downloading  [===>                                     ] (00m:00s)    2.17 MB/s
Extracting   [========>                                ] (00m:00s)        1 / 5
Finished beautifulsoup4                       (00m:00s)              85
KB    521 KB/s
Downloading  [===>                                     ] (00m:00s)    2.17 MB/s
Extracting   [========>                                ] (00m:00s)        1 / 5
Downloading  [===>                                     ] (00m:00s)    2.17 MB/s
Extracting   [========>                                ] (00m:00s)        1 / 5
Finished certifi                              (00m:00s)             150
KB    916 KB/s
Downloading  [===>                                     ] (00m:00s)    2.17 MB/s
Extracting   [========>                                ] (00m:00s)        1 / 5
Downloading  [===>                                     ] (00m:00s)    2.17 MB/s
Extracting   [================>                        ] (00m:00s)        2 / 5
Downloading  [===>                                     ] (00m:00s)    2.17 MB/s
Extracting   [================>                        ] (00m:00s)        2 / 5
Downloading  [===>                                     ] (00m:00s)    2.17 MB/s
Extracting   [================>                        ] (00m:00s)        2 / 5
Downloading  [===>                                     ] (00m:00s)    2.17 MB/s
Extracting   [========================>                ] (00m:00s)        3 / 5
Downloading  [===>                                     ] (00m:00s)    2.17 MB/s
Extracting   [========================>                ] (00m:00s)        3 / 5
Downloading  [===>                                     ] (00m:00s)    2.17 MB/s
Extracting   [================================>        ] (00m:00s)        4 / 5
Downloading  [=========================================] (00m:00s)   18.69 MB/s
Extracting   [================================>        ] (00m:00s)        4 / 5
```

```
Finished openssl                              (00m:00s)               4
MB     17 MB/s
Downloading  [=========================================] (00m:00s)   18.69 MB/s
Extracting   [================================>        ] (00m:00s)        4 / 5
Downloading  [=========================================] (00m:00s)   18.69 MB/s
Extracting   [================================>        ] (00m:00s)        4 / 5
Downloading  [=========================================] (00m:00s)   18.69 MB/s
Extracting   [================================>        ] (00m:00s)        4 / 5
Downloading  [=========================================] (00m:00s)   18.69 MB/s
Extracting   [=========================================] (00m:00s)        5 / 5
Preparing transaction: done
Verifying transaction: done
Executing transaction: done

                  __    __    __    __
                 /  \  /  \  /  \  /  \
                /    \/    \/    \/    \
        /  / /  / /  / /  /
              /  / \   / \   / \   / \  \____
             /  /   \_/   \_/   \_/   \    o \__,
            / _/                       \_____/  `
            |/




        mamba (0.15.3) supported by @QuantStack

        GitHub:  https://github.com/mamba-org/mamba
        Twitter: https://twitter.com/QuantStack




Looking for: ['html5lib==1.1']

pkgs/main/linux-64       Using cache
pkgs/main/noarch         Using cache
pkgs/r/linux-64          Using cache
pkgs/r/noarch            Using cache

Pinned packages:
  - python 3.7.*
```

```
Transaction

  Prefix: /home/jupyterlab/conda/envs/python

  Updating specs:

   - html5lib==1.1
   - ca-certificates
   - certifi
   - openssl


  Package         Version  Build         Channel                  Size

  Install:


  + html5lib          1.1  pyhd3eb1b0_0  pkgs/main/noarch        91 KB
  + webencodings    0.5.1  py37_1        pkgs/main/linux-64      19 KB

  Summary:

  Install: 2 packages

  Total download: 110 KB


Downloading  [=================================>       ] (00m:00s)  622.07 KB/s
Extracting   [>                                                      ] (--:--)
Finished html5lib                             (00m:00s)               91
KB    622 KB/s
Downloading  [=================================>       ] (00m:00s)  622.07 KB/s
Extracting   [>                                                      ] (--:--)
Downloading  [=================================>       ] (00m:00s)  622.07 KB/s
Extracting   [>                                                      ] (--:--)
Downloading  [=================================>       ] (00m:00s)  622.07 KB/s
Extracting   [>                                                      ] (--:--)
Downloading  [=================================>       ] (00m:00s)  622.07 KB/s
Extracting   [====================>                    ] (00m:00s)        1 / 2
Downloading  [=========================================] (00m:00s)  432.74 KB/s
Extracting   [====================>                    ] (00m:00s)        1 / 2
Finished webencodings                         (00m:00s)               19
KB     75 KB/s
Downloading  [=========================================] (00m:00s)  432.74 KB/s
Extracting   [====================>                    ] (00m:00s)        1 / 2
Downloading  [=========================================] (00m:00s)  432.74 KB/s
Extracting   [====================>                    ] (00m:00s)        1 / 2
```

```
Downloading  [=========================================] (00m:00s)  432.74 KB/s
Extracting   [====================>                    ] (00m:00s)        1 / 2
Downloading  [=========================================] (00m:00s)  432.74 KB/s
Extracting   [=========================================] (00m:00s)        2 / 2
Preparing transaction: done
Verifying transaction: done
Executing transaction: done
Collecting lxml==4.6.4
  Downloading lxml-4.6.4-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.m
anylinux_2_24_x86_64.whl (6.3 MB)
                           6.3/6.3 MB
76.7 MB/s eta 0:00:00:00:0100:01
Installing collected packages: lxml
  Attempting uninstall: lxml
    Found existing installation: lxml 4.9.1
    Uninstalling lxml-4.9.1:
      Successfully uninstalled lxml-4.9.1
ERROR: pip's dependency resolver does not currently take into account all
the packages that are installed. This behaviour is the source of the following
dependency conflicts.
yfinance 0.2.4 requires beautifulsoup4>=4.11.1, but you have beautifulsoup4
4.10.0 which is incompatible.
yfinance 0.2.4 requires lxml>=4.9.1, but you have lxml 4.6.4 which is
incompatible.
Successfully installed lxml-4.6.4
```

[2]:

```python
import pandas as pd
import requests
from bs4 import BeautifulSoup
```

0.1 Using Webscraping to Extract Stock Data Example

First we must use the `request` library to downlaod the webpage, and extract the text. We will extract Netflix stock data https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-PY0220EN-SkillsNetwork/labs/project/netflix_data_webpage.html.

[3]:

```python
url = "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/
 ↪IBMDeveloperSkillsNetwork-PY0220EN-SkillsNetwork/labs/project/
 ↪netflix_data_webpage.html"

data  = requests.get(url).text
```

Next we must parse the text into html using `beautiful_soup`

[4]:

```
soup = BeautifulSoup(data, 'html5lib')
```

Now we can turn the html table into a pandas dataframe

[5]:

```
netflix_data = pd.DataFrame(columns=["Date", "Open", "High", "Low", "Close",␣
 ↪"Volume"])

# First we isolate the body of the table which contains all the information
# Then we loop through each row and find all the column values for each row
for row in soup.find("tbody").find_all('tr'):
    col = row.find_all("td")
    date = col[0].text
    Open = col[1].text
    high = col[2].text
    low = col[3].text
    close = col[4].text
    adj_close = col[5].text
    volume = col[6].text

    # Finally we append the data of each row to the table
    netflix_data = netflix_data.append({"Date":date, "Open":Open, "High":high,␣
 ↪"Low":low, "Close":close, "Adj Close":adj_close, "Volume":volume},␣
 ↪ignore_index=True)
```

We can now print out the dataframe

[6]:

```
netflix_data.head()
```

[6]:

```
           Date    Open    High     Low   Close       Volume Adj Close
0  Jun 01, 2021  504.01  536.13  482.14  528.21   78,560,600    528.21
1  May 01, 2021  512.65  518.95  478.54  502.81   66,927,600    502.81
2  Apr 01, 2021  529.93  563.56  499.00  513.47  111,573,300    513.47
3  Mar 01, 2021  545.57  556.99  492.85  521.66   90,183,900    521.66
4  Feb 01, 2021  536.79  566.65  518.28  538.85   61,902,300    538.85
```

We can also use the pandas `read_html` function using the url

[7]:

```
read_html_pandas_data = pd.read_html(url)
```

Or we can convert the BeautifulSoup object to a string

[8]:

```
read_html_pandas_data = pd.read_html(str(soup))
```

Beacause there is only one table on the page, we just take the first table in the list returned

[9]:

```
netflix_dataframe = read_html_pandas_data[0]

netflix_dataframe.head()
```

```
[9]:            Date    Open    High     Low  Close* Adj Close**     Volume
     0  Jun 01, 2021  504.01  536.13  482.14  528.21      528.21   78560600
     1  May 01, 2021  512.65  518.95  478.54  502.81      502.81   66927600
     2  Apr 01, 2021  529.93  563.56  499.00  513.47      513.47  111573300
     3  Mar 01, 2021  545.57  556.99  492.85  521.66      521.66   90183900
     4  Feb 01, 2021  536.79  566.65  518.28  538.85      538.85   61902300
```

0.2 Using Webscraping to Extract Stock Data Exercise

Use the requests library to download the webpage https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-PY0220EN-SkillsNetwork/labs/project/amazon_data_webpage.html. Save the text of the response as a variable named html_data.

```
[10]: url = "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/
        ↪IBMDeveloperSkillsNetwork-PY0220EN-SkillsNetwork/labs/project/
        ↪amazon_data_webpage.html"

      data  = requests.get(url).text
```

Parse the html data using beautiful_soup.

```
[11]: soup = BeautifulSoup(data, 'html5lib')
```

Question 1 What is the content of the title attribute:

```
[20]: soup.title
```

```
[20]: <title>Amazon.com, Inc. (AMZN) Stock Historical Prices &amp; Data - Yahoo
      Finance</title>
```

Using beautiful soup extract the table with historical share prices and store it into a dataframe named amazon_data. The dataframe should have columns Date, Open, High, Low, Close, Adj Close, and Volume. Fill in each variable with the correct data from the list col.

```
[13]: amazon_data = pd.DataFrame(columns=["Date", "Open", "High", "Low", "Close",␣
        ↪"Volume"])

      for row in soup.find("tbody").find_all('tr'):
          col = row.find_all("td")
          date = col[0].text
          Open = col[1].text
          high = col[2].text
          low = col[3].text
          close = col[4].text
          adj_close = col[5].text
          volume = col[6].text
```

```
    amazon_data = amazon_data.append({"Date":date, "Open":Open, "High":high,␣
↪"Low":low, "Close":close, "Adj Close":adj_close, "Volume":volume},␣
↪ignore_index=True)
```

Print out the first five rows of the `amazon_data` dataframe you created.

[14]:
```
amazon_data.head()
```

[14]:
```
           Date      Open      High       Low     Close       Volume Adj Close
0  Jan 01, 2021  3,270.00  3,363.89  3,086.00  3,206.20   71,528,900  3,206.20
1  Dec 01, 2020  3,188.50  3,350.65  3,072.82  3,256.93   77,556,200  3,256.93
2  Nov 01, 2020  3,061.74  3,366.80  2,950.12  3,168.04   90,810,500  3,168.04
3  Oct 01, 2020  3,208.00  3,496.24  3,019.00  3,036.15  116,226,100  3,036.15
4  Sep 01, 2020  3,489.58  3,552.25  2,871.00  3,148.73  115,899,300  3,148.73
```

Question 2 What is the name of the columns of the dataframe

[16]:
```
for col in amazon_data.columns:
    print(col)
```

```
Date
Open
High
Low
Close
Volume
Adj Close
```

Question 3 What is the `Open` of the last row of the amazon_data dataframe?

[15]:
```
amazon_data.tail()
```

[15]:
```
            Date    Open    High     Low   Close       Volume Adj Close
56  May 01, 2016  663.92  724.23  656.00  722.79   90,614,500    722.79
57  Apr 01, 2016  590.49  669.98  585.25  659.59   78,464,200    659.59
58  Mar 01, 2016  556.29  603.24  538.58  593.64   94,009,500    593.64
59  Feb 01, 2016  578.15  581.80  474.00  552.52  124,144,800    552.52
60  Jan 01, 2016  656.29  657.72  547.18  587.00  130,200,900    587.00
```

About the Authors:

Joseph Santarcangelo has a PhD in Electrical Engineering, his research focused on using machine learning, signal processing, and computer vision to determine how videos impact human cognition. Joseph has been working for IBM since he completed his PhD.

Azim Hirjani

0.3 Change Log

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2021-06-09	1.2	Lakshmi Holla	Added URL in question 3
2020-11-10	1.1	Malika Singla	Deleted the Optional part
2020-08-27	1.0	Malika Singla	Added lab to GitLab

##