Activity 7: Reading Tabular Data from a Web Page and Creating DataFrames

1. Open the page in a separate Chrome/Firefox tab and use something like an Inspect Element tool to view the source HTML and understand its structure
2. Read the page using bs4
3. Find the table structure you will need to deal with (how many tables there are?)
4. Find the right table using bs4
5. Separate the source names and their corresponding data
6. Get the source names from the list of sources you have created
7. Separate the header and data from the data that you separated before for the first source only, and then create a DataFrame using that
8. Repeat the last task for the other two data sources

In [80]:
```
from bs4 import BeautifulSoup
import requests
import pandas as pd
import lxml
```

1. Open the page in a separate Chrome/Firefox tab and use something like an Inspect Element tool to view the source HTML and understand its structure

In [124]:
```
url = "https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)"
r = requests.get(url)
# Using r.content or r.text generates FileNotFoundError: [Errno 2] No suc
h file or directory: b'\n<!DOCTYPE html>..
wikipedia_Data = r.content
wikipedia_Data = "List of countries by GDP (nominal) - Wikipedia.htm"
```

1. Read the page using bs4

In [125]:
```
with open(wikipedia_Data, "r",encoding="utf-8") as fd:
    soup = BeautifulSoup(fd)
fd.close()
```

1. Find the table structure you will need to deal with (how many tables there are?)

In [126]:
```
# The table structure in HTML is as follows:
#<table>
#  <tr>
#    <td>Cell A</td>
#    <td>Cell B</td>
#  </tr>
#</table>
# We'll be looking at the <td> tags. These tags will be the source of dat
a

all_tables = soup.find_all("table")
print("Total number of tables are {} ".format(len(all_tables)))
```

```
Total number of tables are 9
```

1. Find the right table using bs4

In [127]:
```
# Tabless are in 'table' tag
data_table = soup.find("table", {"class": '"wikitable"|}'})
print(type(data_table))
```

```
<class 'bs4.element.Tag'>
```

1. Separate the source names and their corresponding data

In [128]:
```
sources = data_table.tbody.findAll('tr', recursive=False)[0]
sources_list = [td for td in sources.findAll('td')]


print("there are {} sources".format(len(sources_list)))
```

```
there are 3 sources
```

In [129]:
```
data = data_table.tbody.findAll('tr', recursive=False)[1].findAll('td', r
ecursive=False)
```

In [130]:
```
data_tables = []
for td in data:
    data_tables.append(td.findAll('table'))

len(data_tables)
```

Out[130]:
```
3
```

1. Get the source names from the list of sources you have created

In [131]:
```
source_names = [source.findAll('a')[0].getText() for source in sources_list]
print(source_names)
```

```
['International Monetary Fund', 'World Bank', 'United Nations']
```

1. Separate the header and data from the data that you separated before for the first source only, and then create a DataFrame using that

In [132]:
```
header1 = [th.getText().strip() for th in data_tables[0][0].findAll('thead')[0].findAll('th')]
print(header1)

rows1 = data_tables[0][0].findAll('tbody')[0].findAll('tr')[1:]
data_rows1 = [[td.get_text().strip() for td in tr.findAll('td')] for tr in rows1]
df1 = pd.DataFrame(data_rows1, columns=header1)
df1.head()
```

```
['Rank', 'Country', 'GDP(US$MM)']
```

Out[132]:

	Rank	Country	GDP(US$MM)
0	1	United States	19,390,600
1	2	China[n 1]	12,014,610
2	3	Japan	4,872,135
3	4	Germany	3,684,816
4	5	United Kingdom	2,624,529

1. Repeat the last task for the other two data sources

In [133]:
```
# This function finds the text in <td> tag.
def find_right_text(i, td):
    if i == 0:
        return td.getText().strip()
    elif i == 1:
        return td.getText().strip()
    else:
        index = td.text.find("♠")
        return td.text[index+1:].strip()
```

In [134]:
```
header2 = [th.getText().strip() for th in data_tables[1][0].findAll('thea
d')[0].findAll('th')]
print(header2)
rows2 = data_tables[1][0].findAll('tbody')[0].findAll('tr')[1:]


data_rows2 = [[find_right_text(i, td) for i, td in enumerate(tr.findAll(
'td'))] for tr in rows2]
df2 = pd.DataFrame(data_rows2, columns=header2)
df2.head()
```

```
['Rank', 'Country', 'GDP(US$MM)']
```

Out[134]:

	Rank	Country	GDP(US$MM)
0	1	United States	19,390,604
1		European Union[23]	17,277,698
2	2	China[n 4]	12,237,700
3	3	Japan	4,872,137
4	4	Germany	3,677,439

In [135]:
```
header3 = [th.getText().strip() for th in data_tables[2][0].findAll('thea
d')[0].findAll('th')]
rows3 = data_tables[2][0].findAll('tbody')[0].findAll('tr')[1:]

data_rows3 = [[find_right_text(i, td) for i, td in enumerate(tr.findAll(
'td'))] for tr in rows2]
df3 = pd.DataFrame(data_rows3, columns=header3)
df3.head()
```

Out[135]:

	Rank	Country	GDP(US$MM)
0	1	United States	19,390,604
1		European Union[23]	17,277,698
2	2	China[n 4]	12,237,700
3	3	Japan	4,872,137
4	4	Germany	3,677,439