In [1]:

```
import bs4
import requests
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In [2]:

```
request_data=requests.get("https://books.toscrape.com/catalogue/page-1.html")
```

In [3]:

```
print(request_data)
```

```
<Response [200]>
```

In [5]:

```
soup=bs4.BeautifulSoup(request_data.text)
```

In [6]:

```
soup
```

Out[6]:

```
<!DOCTYPE html>
<!--[if lt IE 7]>      <html lang="en-us" class="no-js lt-ie9 lt-ie8 lt
-ie7"> <![endif]--><!--[if IE 7]>         <html lang="en-us" class="no-
js lt-ie9 lt-ie8"> <![endif]--><!--[if IE 8]>         <html lang="en-u
s" class="no-js lt-ie9"> <![endif]--><!--[if gt IE 8]><!--><html class
="no-js" lang="en-us"> <!--<![endif]-->
<head>
<title>
    All products | Books to Scrape - Sandbox
</title>
<meta content="text/html; charset=utf-8" http-equiv="content-type"/>
<meta content="24th Jun 2016 09:30" name="created"/>
<meta content="" name="description"/>
<meta content="width=device-width" name="viewport"/>
<meta content="NOARCHIVE,NOCACHE" name="robots"/>
<!-- Le HTML5 shim, for IE6-8 support of HTML elements -->
<!--[if lt IE 9]>
        <script src="//html5shim.googlecode.com/svn/trunk/html5.js"></s
```

In [7]:

```
soup.select('article li')
```

Out[7]:

```
[]
```

In [8]:

```
soup.select('li')
```

Out[8]:

```
[<li>
 <a href="../index.html">Home</a>
 </li>,
 <li class="active">All products</li>,
 <li>
 <a href="category/books_1/index.html">

                                Books

                        </a>
 <ul>
 <li>
 <a href="category/books/travel_2/index.html">

                                Travel

                        </a>
 </li>
```

In [9]:

```
soup.select('li a')
```

Out[9]:

```
[<a href="../index.html">Home</a>,
 <a href="category/books_1/index.html">

                                Books

                        </a>,
 <a href="category/books/travel_2/index.html">

                                Travel

                        </a>,
 <a href="category/books/mystery_3/index.html">

                                Mystery

                        </a>,
 <a href="category/books/historical-fiction_4/index.html">
```

In [10]:

```
soup.select('li a title')
```

Out[10]:

```
[]
```

In [11]:

```
soup.select('li div p')
```

Out[11]:

```
[<p class="price_color">Â£51.77</p>,
 <p class="instock availability">
 <i class="icon-ok"></i>

         In stock

 </p>,
 <p class="price_color">Â£53.74</p>,
 <p class="instock availability">
 <i class="icon-ok"></i>

         In stock

 </p>,
 <p class="price_color">Â£50.10</p>,
 <p class="instock availability">
 <i class="icon-ok"></i>
```

In [12]:

```
soup.select('article h3 a')
```

Out[12]:

```
[<a href="a-light-in-the-attic_1000/index.html" title="A Light in the Atti
c">A Light in the ...</a>,
 <a href="tipping-the-velvet_999/index.html" title="Tipping the Velvet">Ti
pping the Velvet</a>,
 <a href="soumission_998/index.html" title="Soumission">Soumission</a>,
 <a href="sharp-objects_997/index.html" title="Sharp Objects">Sharp Object
s</a>,
 <a href="sapiens-a-brief-history-of-humankind_996/index.html" title="Sapi
ens: A Brief History of Humankind">Sapiens: A Brief History ...</a>,
 <a href="the-requiem-red_995/index.html" title="The Requiem Red">The Requ
iem Red</a>,
 <a href="the-dirty-little-secrets-of-getting-your-dream-job_994/index.htm
l" title="The Dirty Little Secrets of Getting Your Dream Job">The Dirty Li
ttle Secrets ...</a>,
 <a href="the-coming-woman-a-novel-based-on-the-life-of-the-infamous-femin
ist-victoria-woodhull_993/index.html" title="The Coming Woman: A Novel Bas
ed on the Life of the Infamous Feminist, Victoria Woodhull">The Coming Wom
an: A ...</a>,
 <a href="the-boys-in-the-boat-nine-americans-and-their-epic-quest-for-gol
d-at-the-1936-berlin-olympics_992/index.html" title="The Boys in the Boat:
Nine Americans and Their Epic Quest for Gold at the 1936 Berlin Olympics">
The Boys in the ...</a>,
 <a href="the-black-maria_991/index.html" title="The Black Maria">The Blac
k Maria</a>,
 <a href="starving-hearts-triangular-trade-trilogy-1_990/index.html" title
="Starving Hearts (Triangular Trade Trilogy, #1)">Starving Hearts (Triangu
lar Trade ...</a>,
 <a href="shakespeares-sonnets_989/index.html" title="Shakespeare's Sonnet
s">Shakespeare's Sonnets</a>,
 <a href="set-me-free_988/index.html" title="Set Me Free">Set Me Free</a>,
 <a href="scott-pilgrims-precious-little-life-scott-pilgrim-1_987/index.ht
ml" title="Scott Pilgrim's Precious Little Life (Scott Pilgrim #1)">Scott
Pilgrim's Precious Little ...</a>,
 <a href="rip-it-up-and-start-again_986/index.html" title="Rip it Up and S
tart Again">Rip it Up and ...</a>,
 <a href="our-band-could-be-your-life-scenes-from-the-american-indie-under
ground-1981-1991_985/index.html" title="Our Band Could Be Your Life: Scene
s from the American Indie Underground, 1981-1991">Our Band Could Be ...</a
>,
 <a href="olio_984/index.html" title="Olio">Olio</a>,
 <a href="mesaerion-the-best-science-fiction-stories-1800-1849_983/index.h
tml" title="Mesaerion: The Best Science Fiction Stories 1800-1849">Mesaeri
on: The Best Science ...</a>,
 <a href="libertarianism-for-beginners_982/index.html" title="Libertariani
sm for Beginners">Libertarianism for Beginners</a>,
 <a href="its-only-the-himalayas_981/index.html" title="It's Only the Hima
layas">It's Only the Himalayas</a>]
```

In [13]:

```
soup.select('article div a')
```

Out[13]:

```
[<a href="a-light-in-the-attic_1000/index.html"><img alt="A Light in the A
ttic" class="thumbnail" src="../media/cache/2c/da/2cdad67c44b002e7ead0cc35
693c0e8b.jpg"/></a>,
 <a href="tipping-the-velvet_999/index.html"><img alt="Tipping the Velvet"
class="thumbnail" src="../media/cache/26/0c/260c6ae16bce31c8f8c95daddd9f4a
1c.jpg"/></a>,
 <a href="soumission_998/index.html"><img alt="Soumission" class="thumbnai
l" src="../media/cache/3e/ef/3eef99c9d9adef34639f510662022830.jpg"/></a>,
 <a href="sharp-objects_997/index.html"><img alt="Sharp Objects" class="th
umbnail" src="../media/cache/32/51/3251cf3a3412f53f339e42cac2134093.jpg"/>
</a>,
 <a href="sapiens-a-brief-history-of-humankind_996/index.html"><img alt="S
apiens: A Brief History of Humankind" class="thumbnail" src="../media/cach
e/be/a5/bea5697f2534a2f86a3ef27b5a8c12a6.jpg"/></a>,
 <a href="the-requiem-red_995/index.html"><img alt="The Requiem Red" class
="thumbnail" src="../media/cache/68/33/68339b4c9bc034267e1da611ab3b34f8.jp
g"/></a>,
 <a href="the-dirty-little-secrets-of-getting-your-dream-job_994/index.htm
l"><img alt="The Dirty Little Secrets of Getting Your Dream Job" class="th
umbnail" src="../media/cache/92/27/92274a95b7c251fea59a2b8a78275ab4.jpg"/>
</a>,
 <a href="the-coming-woman-a-novel-based-on-the-life-of-the-infamous-femin
ist-victoria-woodhull_993/index.html"><img alt="The Coming Woman: A Novel
Based on the Life of the Infamous Feminist, Victoria Woodhull" class="thum
bnail" src="../media/cache/3d/54/3d54940e57e662c4dd1f3ff00c78cc64.jpg"/></
a>,
 <a href="the-boys-in-the-boat-nine-americans-and-their-epic-quest-for-gol
d-at-the-1936-berlin-olympics_992/index.html"><img alt="The Boys in the Bo
at: Nine Americans and Their Epic Quest for Gold at the 1936 Berlin Olympi
cs" class="thumbnail" src="../media/cache/66/88/66883b91f6804b2323c8369331
cb7dd1.jpg"/></a>,
 <a href="the-black-maria_991/index.html"><img alt="The Black Maria" class
="thumbnail" src="../media/cache/58/46/5846057e28022268153beff6d352b06c.jp
g"/></a>,
 <a href="starving-hearts-triangular-trade-trilogy-1_990/index.html"><img
alt="Starving Hearts (Triangular Trade Trilogy, #1)" class="thumbnail" src
="../media/cache/be/f4/bef44da28c98f905a3ebec0b87be8530.jpg"/></a>,
 <a href="shakespeares-sonnets_989/index.html"><img alt="Shakespeare's Son
nets" class="thumbnail" src="../media/cache/10/48/1048f63d3b5061cd2f424d20
b3f9b666.jpg"/></a>,
 <a href="set-me-free_988/index.html"><img alt="Set Me Free" class="thumbn
ail" src="../media/cache/5b/88/5b88c52633f53cacf162c15f4f823153.jpg"/></a
>,
 <a href="scott-pilgrims-precious-little-life-scott-pilgrim-1_987/index.ht
ml"><img alt="Scott Pilgrim's Precious Little Life (Scott Pilgrim #1)" cla
ss="thumbnail" src="../media/cache/94/b1/94b1b8b244bce9677c2f29ccc890d4d2.
jpg"/></a>,
 <a href="rip-it-up-and-start-again_986/index.html"><img alt="Rip it Up an
d Start Again" class="thumbnail" src="../media/cache/81/c4/81c4a973364e17d
01f217e1188253d5e.jpg"/></a>,
 <a href="our-band-could-be-your-life-scenes-from-the-american-indie-under
ground-1981-1991_985/index.html"><img alt="Our Band Could Be Your Life: Sc
enes from the American Indie Underground, 1981-1991" class="thumbnail" src
="../media/cache/54/60/54607fe8945897cdcced0044103b10b6.jpg"/></a>,
 <a href="olio_984/index.html"><img alt="Olio" class="thumbnail" src="../m
edia/cache/55/33/553310a7162dfbc2c6d19a84da0df9e1.jpg"/></a>,
 <a href="mesaerion-the-best-science-fiction-stories-1800-1849_983/index.h
tml"><img alt="Mesaerion: The Best Science Fiction Stories 1800-1849" clas
s="thumbnail" src="../media/cache/09/a3/09a3aef48557576e1a85ba7efea8ecb7.j
pg"/></a>,
 <a href="libertarianism-for-beginners_982/index.html"><img alt="Libertari
```

In [14]:
```
len(soup.select('article div a'))
```

Out[14]:
```
20
```

In [16]:
```
soup.select('article div')[0]['class']
```

Out[16]:
```
['image_container']
```

In [18]:
```
title=[]
rating=[]
for i in range(1,51):
    req_page=requests.get(f'https://books.toscrape.com/catalogue/page-{i}.html')
    soup_page=bs4.BeautifulSoup(req_page.text)
    for j in soup_page.select('article h3 a'):
        title.append(j['title'])
    for k in range(0,60,3):
        rating.append(soup_page.select('article p')[k]['class'][1])
```

```
anism for Beginners" class="thumbnail" src="../media/cache/0b/bc/0bbcd0a6f
4bcd81ccb1049a52736406e.jpg"/></a>,
 <a href="its-only-the-himalayas_981/index.html"><img alt="It's Only the H
imalayas" class="thumbnail" src="../media/cache/27/a5/27a53d0bb95bdd88288e
af66c9230d7e.jpg"/></a>]
```

In [20]:

```
for i in title:
    print(i)
```

```
A Light in the Attic
Tipping the Velvet
Soumission
Sharp Objects
Sapiens: A Brief History of Humankind
The Requiem Red
The Dirty Little Secrets of Getting Your Dream Job
The Coming Woman: A Novel Based on the Life of the Infamous Feminist, V
ictoria Woodhull
The Boys in the Boat: Nine Americans and Their Epic Quest for Gold at t
he 1936 Berlin Olympics
The Black Maria
Starving Hearts (Triangular Trade Trilogy, #1)
Shakespeare's Sonnets
Set Me Free
Scott Pilgrim's Precious Little Life (Scott Pilgrim #1)
Rip it Up and Start Again
Our Band Could Be Your Life: Scenes from the American Indie Undergroun
d, 1981-1991
```

In [21]:

```
for i in rating:
    print(i)
```

```
Three
One
One
Four
Five
One
Four
Three
Four
One
Two
Four
Five
Five
Five
Three
One
One
Two
```

In [26]:

```
data=dict(zip(title,rating))
```

In [27]:

```
data
```

Out[27]:


```
{'A Light in the Attic': 'Three',
 'Tipping the Velvet': 'One',
 'Soumission': 'One',
 'Sharp Objects': 'Four',
 'Sapiens: A Brief History of Humankind': 'Five',
 'The Requiem Red': 'One',
 'The Dirty Little Secrets of Getting Your Dream Job': 'Four',
 'The Coming Woman: A Novel Based on the Life of the Infamous Feminist,
Victoria Woodhull': 'Three',
 'The Boys in the Boat: Nine Americans and Their Epic Quest for Gold at
the 1936 Berlin Olympics': 'Four',
 'The Black Maria': 'One',
 'Starving Hearts (Triangular Trade Trilogy, #1)': 'Two',
 "Shakespeare's Sonnets": 'Four',
 'Set Me Free': 'Five',
 "Scott Pilgrim's Precious Little Life (Scott Pilgrim #1)": 'Five',
 'Rip it Up and Start Again': 'Five',
 'Our Band Could Be Your Life: Scenes from the American Indie Undergrou
```


In [28]:

```
dataset=pd.DataFrame.from_dict(dataset,orient='index')
```

In [29]:

```
dataset
```

Out[29]:

	0
A Light in the Attic	Three
Tipping the Velvet	One
Soumission	One
Sharp Objects	Four
Sapiens: A Brief History of Humankind	Five
...	...
Alice in Wonderland (Alice's Adventures in Wonderland #1)	One
Ajin: Demi-Human, Volume 1 (Ajin: Demi-Human #1)	Four
A Spy's Devotion (The Regency Spies of London #1)	Five
1st to Die (Women's Murder Club #1)	One
1,000 Places to See Before You Die	Five

999 rows × 1 columns

In [30]:

```
dataset.to_csv('scrapped_data.csv')
```

In [31]:

```
pwd
```

Out[31]:

```
'C:\\Users\\Lenovo'
```

In [32]:

```
import os
cd=os.getcwd()
print(cd)
```

```
C:\Users\Lenovo
```

In [33]:

```
new_books=pd.read_csv('scrapped_data.csv')
```

In [34]:

```
new_books
```

Out[34]:

	Unnamed: 0	0
0	A Light in the Attic	Three
1	Tipping the Velvet	One
2	Soumission	One
3	Sharp Objects	Four
4	Sapiens: A Brief History of Humankind	Five
...	...	...
994	Alice in Wonderland (Alice's Adventures in Won...	One
995	Ajin: Demi-Human, Volume 1 (Ajin: Demi-Human #1)	Four
996	A Spy's Devotion (The Regency Spies of London #1)	Five
997	1st to Die (Women's Murder Club #1)	One
998	1,000 Places to See Before You Die	Five

999 rows × 2 columns

In [35]:

```
soup.select('article h3 a')[0]['href']
```

Out[35]:

```
'a-light-in-the-attic_1000/index.html'
```

In [36]:

```
soup.select('tr')
```

Out[36]:

```
[]
```

In [37]:

```
new_request = requests.get('https://books.toscrape.com/catalogue/in-her-wake_980/index.h
```

In [38]:

```
new_request
```

Out[38]:

```
<Response [200]>
```

In [39]:

```
soup_new=bs4.BeautifulSoup(new_request.text)
```

In [40]:

```
soup_new.select('tr')
```

Out[40]:

```
[<tr>
 <th>UPC</th><td>23356462d1320d61</td>
 </tr>,
 <tr>
 <th>Product Type</th><td>Books</td>
 </tr>,
 <tr>
 <th>Price (excl. tax)</th><td>Â£12.84</td>
 </tr>,
 <tr>
 <th>Price (incl. tax)</th><td>Â£12.84</td>
 </tr>,
 <tr>
 <th>Tax</th><td>Â£0.00</td>
 </tr>,
 <tr>
 <th>Availability</th>
 <td>In stock (19 available)</td>
 </tr>,
 <tr>
 <th>Number of reviews</th>
 <td>0</td>
 </tr>]
```

In [41]:

```
soup_new.select('tr th')[2].text
```

Out[41]:

```
'Price (excl. tax)'
```

In [42]:

```
new_df = pd.DataFrame()  # empty dataframe
index_num= 0             # index of dataframe

for page_num in range(1,51):    # total pages
    request = requests.get(f'http://books.toscrape.com/catalogue/page-{page_num}.html')
    soup_ = bs4.BeautifulSoup(request.text)
    book_info_list = soup_.select('article h3 a')   # all books on that page

    for n in range(0,20):
        res = requests.get('http://books.toscrape.com/catalogue/'+ book_info_list[n]['hr
        soup = bs4.BeautifulSoup(res.text, 'lxml')

        book_features_column_list = soup.select('tr th')   # book features like price, r
        book_column_values = soup.select('tr td')          # values of that features

        column_names = [ item.text for item in book_features_column_list ]

        column_values = [item.text for item in book_column_values ]

        d = dict(zip(column_names, column_values))  # making dictionary with features as

        df = pd.DataFrame(d, index = [index_num + n])
        new_df = new_df.append(df)    # append df of every page

    index_num += 20  # index for each page to be incremented by 20 as zero_base index si
```

```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_12112\3760563315.py:23: Fu
tureWarning: The frame.append method is deprecated and will be removed
from pandas in a future version. Use pandas.concat instead.
  new_df = new_df.append(df)    # append df of every page
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_12112\3760563315.py:23: Fu
tureWarning: The frame.append method is deprecated and will be removed
from pandas in a future version. Use pandas.concat instead.
  new_df = new_df.append(df)    # append df of every page
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_12112\3760563315.py:23: Fu
tureWarning: The frame.append method is deprecated and will be removed
from pandas in a future version. Use pandas.concat instead.
  new_df = new_df.append(df)    # append df of every page
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_12112\3760563315.py:23: Fu
tureWarning: The frame.append method is deprecated and will be removed
from pandas in a future version. Use pandas.concat instead.
  new_df = new_df.append(df)    # append df of every page
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_12112\3760563315.py:23: Fu
tureWarning: The frame.append method is deprecated and will be removed
from pandas in a future version. Use pandas.concat instead.
```

In [ ]:

```
new_df
```

In [ ]:

```

```