Web Scrapping for Book store website

1.Extraction

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

I have imported the `requests` library for making HTTP requests, `BeautifulSoup` from the `bs4` library for web scraping, and `pandas` for data manipulation in a single line using the Python `import` statement.

```
books = []

for i in range(1,20):
  url = f"https://books.toscrape.com/catalogue/page-{i}.html"
  response = requests.get(url)
  response = response.content
  soup = BeautifulSoup(response, 'html.parser')
  ol = soup.find('ol')
  articles = ol.find_all('article', class_='product_pod')
  for article in articles:
    image = article.find('img')
    title = image.attrs['alt']
    starTag = article.find('p')
    star = starTag['class'][1]
    price = article.find('p', class_='price_color').text
    price = float(price[1:])
    books.append([title, star, price])


df = pd.DataFrame(books, columns=['Title', 'Star Rating', 'Price'])
df.to_csv('books.csv')
```

I wrote a Python script to scrape book information from a website. I looped through multiple pages, extracted details like title, star rating, and price of each book, and stored them in a list. Then, I created a DataFrame using the pandas library and saved it as a CSV file named 'books.csv'. This allows me to analyze and work with the data more easily.

```
print(df)
```

```
                                                Title Star Rating  Price
0                                A Light in the Attic       Three  51.77
1                                  Tipping the Velvet         One  53.74
2                                          Soumission         One  50.10
3                                       Sharp Objects        Four  47.82
4               Sapiens: A Brief History of Humankind        Five  54.23
5                                     The Requiem Red         One  22.65
6    The Dirty Little Secrets of Getting Your Dream...        Four  33.34
7    The Coming Woman: A Novel Based on the Life of...       Three  17.93
8    The Boys in the Boat: Nine Americans and Their...        Four  22.60
9                                     The Black Maria         One  52.15
10     Starving Hearts (Triangular Trade Trilogy, #1)         Two  13.99
11                             Shakespeare's Sonnets        Four  20.66
12                                     Set Me Free        Five  17.46
13   Scott Pilgrim's Precious Little Life (Scott Pi...        Five  52.29
14                       Rip it Up and Start Again        Five  35.02
15   Our Band Could Be Your Life: Scenes from the A...       Three  57.25
16                                            Olio         One  23.88
17   Mesaerion: The Best Science Fiction Stories 18...         One  37.59
18                     Libertarianism for Beginners         Two  51.33
19                          It's Only the Himalayas         Two  45.17
20                                     In Her Wake         One  12.84
21                                  How Music Works         Two  37.32
22   Foolproof Preserving: A Guide to Small Batch J...       Three  30.52
23                         Chase Me (Paris Nights #2)        Five  25.27
```

24	Black Dust	Five	34.53
25	Birdsong: A Story in Pictures	Three	54.64
26	America's Cradle of Quarterbacks: Western Penn...	Three	22.50
27	Aladdin and His Wonderful Lamp	Three	53.13
28	Worlds Elsewhere: Journeys Around Shakespeare’...	Five	40.30
29	Wall and Piece	Four	44.18
30	The Four Agreements: A Practical Guide to Pers...	Five	17.66
31	The Five Love Languages: How to Express Heartf...	Three	31.05
32	The Elephant Tree	Five	23.82
33	The Bear and the Piano	One	36.89
34	Sophie's World	Five	15.94
35	Penny Maybe	Three	33.29
36	Maude (1883-1993):She Grew Up with the country	Two	18.02
37	In a Dark, Dark Wood	One	19.63
38	Behind Closed Doors	Four	52.22
39	You can't bury them all: Poems	Two	33.63
40	Slow States of Collapse: Poems	Three	57.31
41	Reasons to Stay Alive	Two	26.41
42	Private Paris (Private #10)	Five	47.61
43	#HigherSelfie: Wake Up Your Life. Free Your So...	Five	23.11
44	Without Borders (Wanderlove #1)	Two	45.07
45	When We Collided	One	31.77
46	We Love You, Charlie Freeman	Five	50.27
47	Untitled Collection: Sabbath Poems 2014	Four	14.27
48	Unseen City: The Majesty of Pigeons, the Discr...	Four	44.18
49	Unicorn Tracks	Three	18.78
50	Unbound: How Eight Technologies Made Us Human,...	One	25.52
51	Tsubasa: WoRLD CHRoNiCLE 2 (Tsubasa WoRLD CHRo...	One	16.28
52	Throwing Rocks at the Google Bus: How Growth B...	Three	31.12
53	This One Summer	Four	19.49
54	Thirst	Five	17.27
55	The Torch Is Passed: A Harding Family Story	One	19.09
56	The Secret of Broadvilla Cross	One	56.13

I tried to print the dataframe but it is printing on first few lines and last lines . so i used display max rows to see all the rows and columns .

```
pd.set_option('display.max_rows', None)
print(df)
```

```
353                                          Rook        Four  37.86
354    My Kitchen Year: 136 Recipes That Saved My Life        Two  11.53
355  13 Hours: The Inside Account of What Really Ha...        One  27.06
356                         Will You Won't You Want Me?      Three  13.86
357  Tipping Point for Planet Earth: How Close Are ...        One  37.55
358                              The Star-Touched Queen       Five  32.30
359            The Silent Sister (Riley MacPherson #1)       Five  46.29
360  The Midnight Watch: A Novel of the Titanic and...        One  26.20
361  The Lonely City: Adventures in the Art of Bein...        Two  33.26
362  The Gray Rhino: How to Recognize and Act on th...       Four  59.15
363  The Golden Condom: And Other Essays on Love Lo...        One  39.43
```

2.Transformation of Data

Since , I am able to sucessfully fetch the data of the columns , Title , Name ,Price and Rating for the Book website . I looked throughly through whole data to make possible transformations for the data.so , I have converted the ratings and price into numerical data inorder to procedd with further numerical operations and analysis.

```
df['Star Rating'] = df['Star Rating'].map({'One': 1, 'Two': 2, 'Three': 3, 'Four': 4, 'Five': 5
df['Price'] = pd.to_numeric(df['Price'])
```

```
print(df)
```

```
322          City of Glass (The Mortal Instruments #3)           4  56.02
323  City of Fallen Angels (The Mortal Instruments #4)           4  11.23
324          City of Bones (The Mortal Instruments #1)           1  43.28
325          City of Ashes (The Mortal Instruments #2)           1  47.27
326                                               Cell           4  20.29
327                      Catching Jordan (Hundred Oaks)           3  50.83
328       Carry On, Warrior: Thoughts on Life Unarmed             3  31.85
329                                             Carrie           2  46.23
330  Buying In: The Secret Dialogue Between What We...           4  37.80
331                 Brain on Fire: My Month of Madness           5  49.32
332                                      Batman: Europa           2  32.01
333                   Barefoot Contessa Back to Basics           1  28.01
334  Barefoot Contessa at Home: Everyday Recipes Yo...           5  50.62
335                                    Balloon Animals           3  17.03
336                                     Art Ops Vol. 1           3  48.80
337  Aristotle and Dante Discover the Secrets of th...           4  58.14
338                 Angels Walking (Angels Walking #1)           2  34.20
339                Angels & Demons (Robert Langdon #1)           3  51.48
340                        All the Light We Cannot See           5  29.87
341  Adulthood Is a Myth: A "Sarah's Scribbles" Col...           2  10.90
342                                      Abstract City           5  56.37
343            A Time of Torment (Charlie Parker #14)           5  48.35
344           A Study in Scarlet (Sherlock Holmes #1)           2  16.73
345  A Series of Catastrophes and Miracles: A True ...           2  56.48
346           A People's History of the United States            2  40.79
347                                    A Man Called Ove           1  39.72
348      A Distant Mirror: The Calamitous 14th Century           3  14.58
349               A Brush of Wings (Angels Walking #3)           1  55.51
350  1491: New Revelations of the Americas Before C...           3  21.80
351         The Three Searches, Meaning, and the Story           3  13.33
352                   Searching for Meaning in Gailana           1  38.73
353                                               Rook           4  37.86
354    My Kitchen Year: 136 Recipes That Saved My Life           2  11.53
355  13 Hours: The Inside Account of What Really Ha...           1  27.06
356                         Will You Won't You Want Me?           3  13.86
357  Tipping Point for Planet Earth: How Close Are ...           1  37.55
358                              The Star-Touched Queen           5  32.30
359            The Silent Sister (Riley MacPherson #1)           5  46.29
360  The Midnight Watch: A Novel of the Titanic and...           1  26.20
361  The Lonely City: Adventures in the Art of Bein...           2  33.26
362  The Gray Rhino: How to Recognize and Act on th...           4  59.15
363  The Golden Condom: And Other Essays on Love Lo...           1  39.43
364                     The Epidemic (The Program 0.6)           5  14.44
365                                  The Dinner Party            2  56.54
366                            The Diary of a Young Girl         3  59.90
367                                       The Children           3  11.88
368            Stars Above (The Lunar Chronicles #4.5)           2  48.05
369  Snatched: How A Drug Queen Went Undercover for...           3  21.21
370  Raspberry Pi Electronics Projects for the Evil...           4  49.67
371  Quench Your Own Thirst: Business Lessons Learn...           1  43.14
372                   Psycho: Sanitarium (Psycho #1.5)           5  36.97
373                   Poisonous (Max Revere Novels #3)           3  26.80
374                        One with You (Crossfire #5)           4  15.71
375                    No Love Allowed (Dodge Cove #1)           4  54.65
376  Murder at the 42nd Street Library (Raymond Amb...           4  54.36
377                                       Most Wanted           3  35.28
378                               Love, Lies and Spies          2  20.55
379  How to Speak Golf: An Illustrated Guide to Lin...           5  58.32
```

I tried to print the most expensive book. i.e whose cost is more than 50

```
expensive_books = df[df['Price'] > 50]

print(expensive_books)
```

```
     61   The Murder That Never Was (Forensic Instincts #5)           3  54.11
     67   The Electric Pencil: Drawings from Inside Stat...           1  56.06
     68         The Death of Humanity: and the Case for Life          4  58.11
     77       Saga, Volume 5 (Saga (Collected Editions) #5)           2  51.04
     79   Rat Queens, Vol. 3: Demons (Rat Queens (Collec...           3  50.40
     91                                  Masks and Shadows            2  56.40
     97   Judo: Seven Steps to Black Belt (an Introducto...           2  53.90
     100  Immunity: How Elie Metchnikoff Changed the Cou...           5  57.36
     102                      I am a Hero Omnibus Volume 1            3  54.63
     109  Everydata: The Misinformation Hidden in the Li...           2  54.35
     111                               Danganronpa Volume 1           4  51.99
     122  A Piece of Sky, a Grain of Rice: A Memoir in F...           5  56.76
     124            A Flight of Arrows (The Pathfinders #2)           5  55.53
     126  A Court of Thorns and Roses (A Court of Thorns...           1  52.37
     127  (Un)Qualified: How God Uses Broken People to D...           5  54.00
     133  Thomas Jefferson and the Tripoli Pirates: The ...           1  59.64
     134                               Thirteen Reasons Why           1  52.72
     135  The White Cat and the Monk: A Retelling of the...           4  58.08
     142               The Regional Office Is Under Attack!           5  51.36
     150       The Matchmaker's Playbook (Wingmen Inc. #1)           1  55.85
     156                               The Girl on the Train          2  55.02
     165                   Suddenly in Love (Lake Haven #1)           2  55.99
     186  I Had a Nice Time And Other Lies...: How to fi...           4  57.36
     193           Finders Keepers (Bill Hodges Trilogy #2)           5  53.53
     195                                 Eureka Trivia 6.0            4  54.59
     213                                             Amatus           5  50.54
     219  Why the Right Went Wrong: Conservatism--From G...           4  52.65
     222  We Are Robin, Vol. 1: The Vigilante Business (...           1  53.90
     231                               The Wright Brothers            4  56.80
     235                              The Testament of Mary           4  52.67
     239                 The Rosie Project (Don Tillman #1)           1  54.04
     246                     The Last Mile (Amos Decker #2)           2  54.21
     248       The Hidden Oracle (The Trials of Apollo #1)           2  52.26
     265                Someone Like You (The Harrisons #2)           5  52.79
     267                                              Shtum           4  55.84
     273  Quarter Life Poetry: Poems for the Young, Brok...           5  50.89
     275  Overload: How to Unplug, Unwind, and Unleash Y...           3  52.15
     286                               Luis Paints the World          3  53.95
     288  Lowriders to the Center of the Earth (Lowrider...           2  51.51
     301                           Hamilton: The Revolution           3  58.79
     309                                            El Deafo          5  57.62
     311  Eaternity: More than 150 Deliciously Easy Vega...           5  51.75
     312                                 Eat Fat, Get Thin            2  54.07
     313                                  Don't Get Caught            1  55.35
     322         City of Glass (The Mortal Instruments #3)           4  56.02
     327                     Catching Jordan (Hundred Oaks)          3  50.83
     334  Barefoot Contessa at Home: Everyday Recipes Yo...           5  50.62
     337  Aristotle and Dante Discover the Secrets of th...           4  58.14
     339                Angels & Demons (Robert Langdon #1)           3  51.48
     342                                      Abstract City          5  56.37
     345  A Series of Catastrophes and Miracles: A True ...           2  56.48
     349               A Brush of Wings (Angels Walking #3)           1  55.51
     362  The Gray Rhino: How to Recognize and Act on th...           4  59.15
     365                                  The Dinner Party            2  56.54
     366                          The Diary of a Young Girl           3  59.90
     375                    No Love Allowed (Dodge Cove #1)           4  54.65
     376  Murder at the 42nd Street Library (Raymond Amb...           4  54.36
     379  How to Speak Golf: An Illustrated Guide to Lin...           5  58.32
```

I used the `sort_values` function on my DataFrame. It rearranged the rows based on the 'Price' column in descending order, meaning the books with the highest prices are listed first. After sorting, I printed the updated DataFrame to see the changes.

```
sorted_df = df.sort_values(by='Price', ascending=False)
print(sorted_df)
```

```
                                            Title  Star Rating  Price
     366                          The Diary of a Young Girl           3  59.90
     133  Thomas Jefferson and the Tripoli Pirates: The ...           1  59.64
     362  The Gray Rhino: How to Recognize and Act on th...           4  59.15
     301                           Hamilton: The Revolution           3  58.79
     379  How to Speak Golf: An Illustrated Guide to Lin...           5  58.32
     337  Aristotle and Dante Discover the Secrets of th...           4  58.14
     68         The Death of Humanity: and the Case for Life          4  58.11
     135  The White Cat and the Monk: A Retelling of the...           4  58.08
     309                                            El Deafo          5  57.62
     100  Immunity: How Elie Metchnikoff Changed the Cou...           5  57.36
     186  I Had a Nice Time And Other Lies...: How to fi...           4  57.36
```

```
40                      Slow States of Collapse: Poems           3  57.31
15   Our Band Could Be Your Life: Scenes from the A...           3  57.25
231                               The Wright Brothers           4  56.80
122  A Piece of Sky, a Grain of Rice: A Memoir in F...           5  56.76
365                                   The Dinner Party           2  56.54
58                                  The Past Never Ends          4  56.50
345  A Series of Catastrophes and Miracles: A True ...           2  56.48
57   The Pioneer Woman Cooks: Dinnertime: Comfort C...           1  56.41
91                                    Masks and Shadows          2  56.40
342                                       Abstract City          5  56.37
56                     The Secret of Dreadwillow Carse           1  56.13
67   The Electric Pencil: Drawings from Inside Stat...           1  56.06
322          City of Glass (The Mortal Instruments #3)           4  56.02
165                   Suddenly in Love (Lake Haven #1)           2  55.99
150        The Matchmaker's Playbook (Wingmen Inc. #1)           1  55.85
267                                               Shtum           4  55.84
124            A Flight of Arrows (The Pathfinders #2)           5  55.53
349               A Brush of Wings (Angels Walking #3)           1  55.51
313                                 Don't Get Caught             1  55.35
156                             The Girl on the Train            2  55.02
375                    No Love Allowed (Dodge Cove #1)           4  54.65
25                       Birdsong: A Story in Pictures           3  54.64
102                       I am a Hero Omnibus Volume 1           3  54.63
195                                 Eureka Trivia 6.0            4  54.59
376  Murder at the 42nd Street Library (Raymond Amb...           4  54.36
109  Everydata: The Misinformation Hidden in the Li...           2  54.35
4                Sapiens: A Brief History of Humankind           5  54.23
246                     The Last Mile (Amos Decker #2)           2  54.21
61   The Murder That Never Was (Forensic Instincts #5)           3  54.11
312                                 Eat Fat, Get Thin            2  54.07
239                 The Rosie Project (Don Tillman #1)           1  54.04
127  (Un)Qualified: How God Uses Broken People to D...           5  54.00
286                               Luis Paints the World          3  53.95
97   Judo: Seven Steps to Black Belt (an Introducto...           2  53.90
222  We Are Robin, Vol. 1: The Vigilante Business (...           1  53.90
1                                   Tipping the Velvet           1  53.74
193           Finders Keepers (Bill Hodges Trilogy #2)           5  53.53
27                       Aladdin and His Wonderful Lamp          3  53.13
265                Someone Like You (The Harrisons #2)           5  52.79
134                               Thirteen Reasons Why           1  52.72
235                               The Testament of Mary          4  52.67
219  Why the Right Went Wrong: Conservatism--From G...           4  52.65
126  A Court of Thorns and Roses (A Court of Thorns...           1  52.37
13   Scott Pilgrim's Precious Little Life (Scott Pi...           5  52.29
248        The Hidden Oracle (The Trials of Apollo #1)           2  52.26
38                                  Behind Closed Doors          4  52.22
```

I used the `groupby` function on my DataFrame, grouping the data by 'Star Rating'. Then, I calculated the average price for each rating category using the 'Price' column. The result, `avg_price_by_rating`, is a Series that shows the mean price for books in each star rating. Printing it provides a quick overview of how prices vary across different ratings.

```
avg_price_by_rating = df.groupby('Star Rating')['Price'].mean()
print(avg_price_by_rating)
```

```
Star Rating
1    34.876585
2    33.656329
3    34.431600
4    36.850145
5    34.315067
Name: Price, dtype: float64
```

I used the `nlargest` function on my DataFrame to find the book with the highest price. The code then prints this costliest book. You can adjust the '1' in `nlargest(1, 'Price')` to display more top costly books if needed. It's a quick way to identify and showcase the most expensive books in the dataset.

```
costliest_books = df.nlargest(1, 'Price')  # Change '1' to the number of top costly books you w
print(costliest_books)
```

```
                         Title  Star Rating  Price
366  The Diary of a Young Girl            3   59.9
```