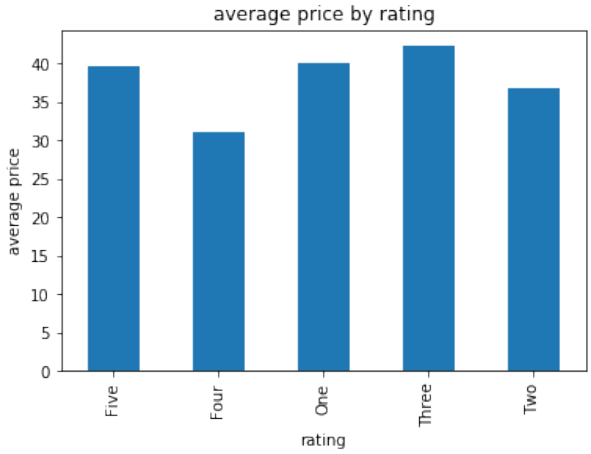
Assignment 3: Web Scraping

Q1. Scrape Book Catalog

- Scape content of http://books.toscrape.com (http://books.toscrape.com)
- Write a function getData() to scrape **title** (see (1) in Figure), **rating** (see (2) in Figure), **price** (see (3) in Figure) of all books (i.e. 20 books) listed in the page.
  - For example, the figure shows one book and the corresponding html code. You need to scrape the highlighted content.
  - For star ratings, you can simply scrape One, Two, Three, ...
  - The highlighted content in the figure should be saved into a tuple ('A Light in the ...','Three','£51.77')
- The output is a list of 20 tuples, e.g. [('A Light in the ...','Three','£51.77'), ...]. Each tuple corresponds to one book.

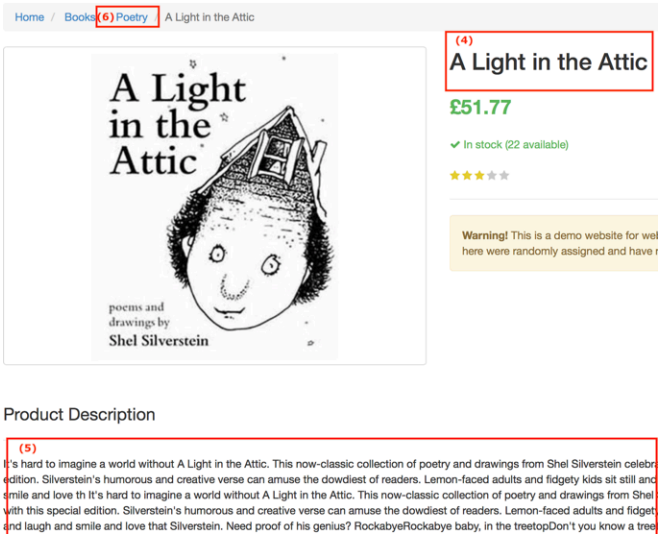
Q2. Plot data

- Create a function plot_data which
  - takes the list of tuples from Q1 as an input
  - converts the price strings to numbers
  - calculates the average price of books by ratings
  - plots a bar chart for the average price. The plot may look similar to the figure below.

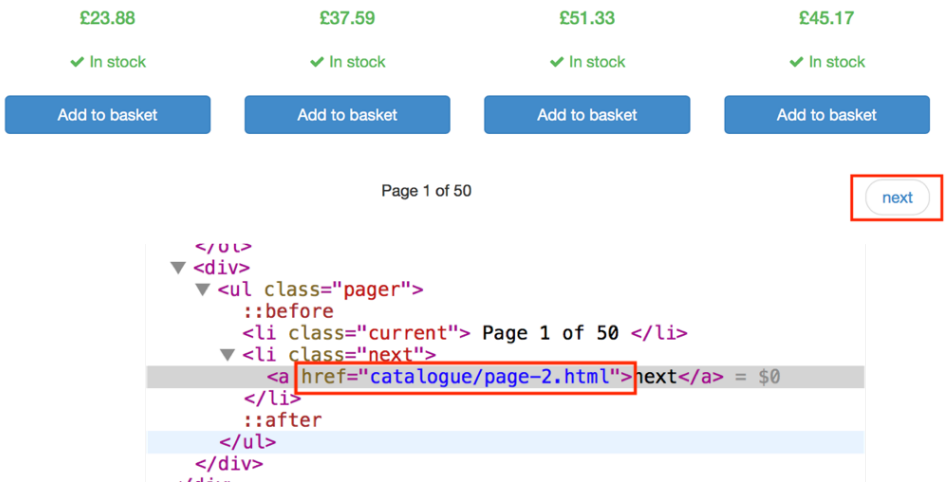


Q3 (Bonus) Expand your solution to Q1 to scrape the full details of all books on http://books.toscrape.com (http://books.toscrape.com)

- Write a function getFullData() to do the following:
  - Besides scraping title, rating, and price of each book as stated in Q1, also scrape the **full title** (see (4) in Figure), **description** (see (5) in Figure), and **category** (see (6) in Figure) in each individual book page.
    - An example individual book page is shown in the figure below.



  - Scape all book listing pages following the "next" link at the bottom. The figure below gives an screenshot of the "next" link and its corresponding html code.
  - Do not hardcode page URLs (except http://books.toscrape.com (http://books.toscrape.com)) in your code.



  - The output is a list containing 1000 tuples,
    - e.g. [('A Light in the ...','Three','£51.77', 'A Light in the Attic', "It's hard to imagine a world without A Light in the Attic. This now-classic collection ...",'Poetry'), ...]

In [5]:
```
import requests
from bs4 import BeautifulSoup
import pandas as pd
import matplotlib.pyplot as plt


 # Q1
def getData():


    data=[]  # variable to hold all book data

    page_url="http://books.toscrape.com"

    # your code here

    return data

#Q2
def plot_data(data):

    # fill your code here


# Q3
def getFullData()
    data=[]

    # fill your code here

    return data



if __name__ == "__main__":

    # Test Q1
    data=getData()
    print(data)

    # Test Q2
    plot_data(data)

    # Test Q3
    data=getFullData()
    print(data)
```

In [ ]: