Web scraping

Presented by: Hameed Emad Hameed And
Ali Khalid Jabbar

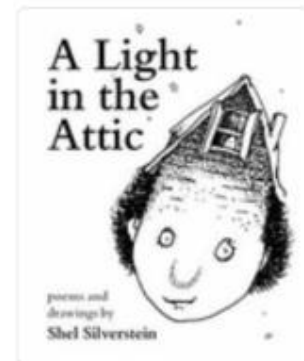UOITC

1

The web we used

2

The code and an explanation of some important points

3
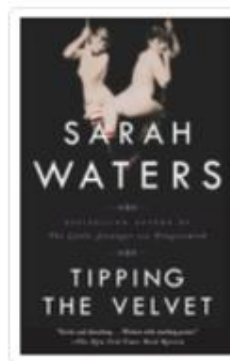
Solve problems and explain the benefit of the project

A Light in the ...

Tipping the Velvet

£51.77

£53.74

✔ In stock

✔ In stock

Add to basket

Add to basket
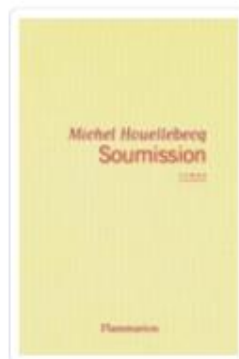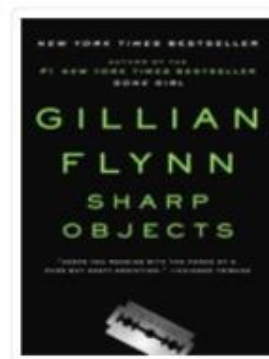
Soumission

Sharp Objects

```html
<div id="messages"> </div>
<div id="promotions"> </div>
<form method="get" class="form-horizontal">⋯</form>
<section>
  <div class="alert alert-warning" role="alert">⋯</div>
  <div>
    <ol class="row">
      ::before
      <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">⋯</li>
      <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">⋯</li>
      <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">⋯</li>
      <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
        <article class="product_pod">
          <div class="image_container">
            <a href="catalogue/sharp-objects_997/index.html">⋯</a> == $0
          </div>
          <p class="star-rating Four">⋯</p>
          <h3>⋯</h3>
          <div class="product_price">⋯</div>
        </article>
      </li>
      <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">⋯</li>
      <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">⋯</li>
      <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">⋯</li>
      <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">⋯</li>
      <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">⋯</li>
      <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">⋯</li>
      <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">⋯</li>
```
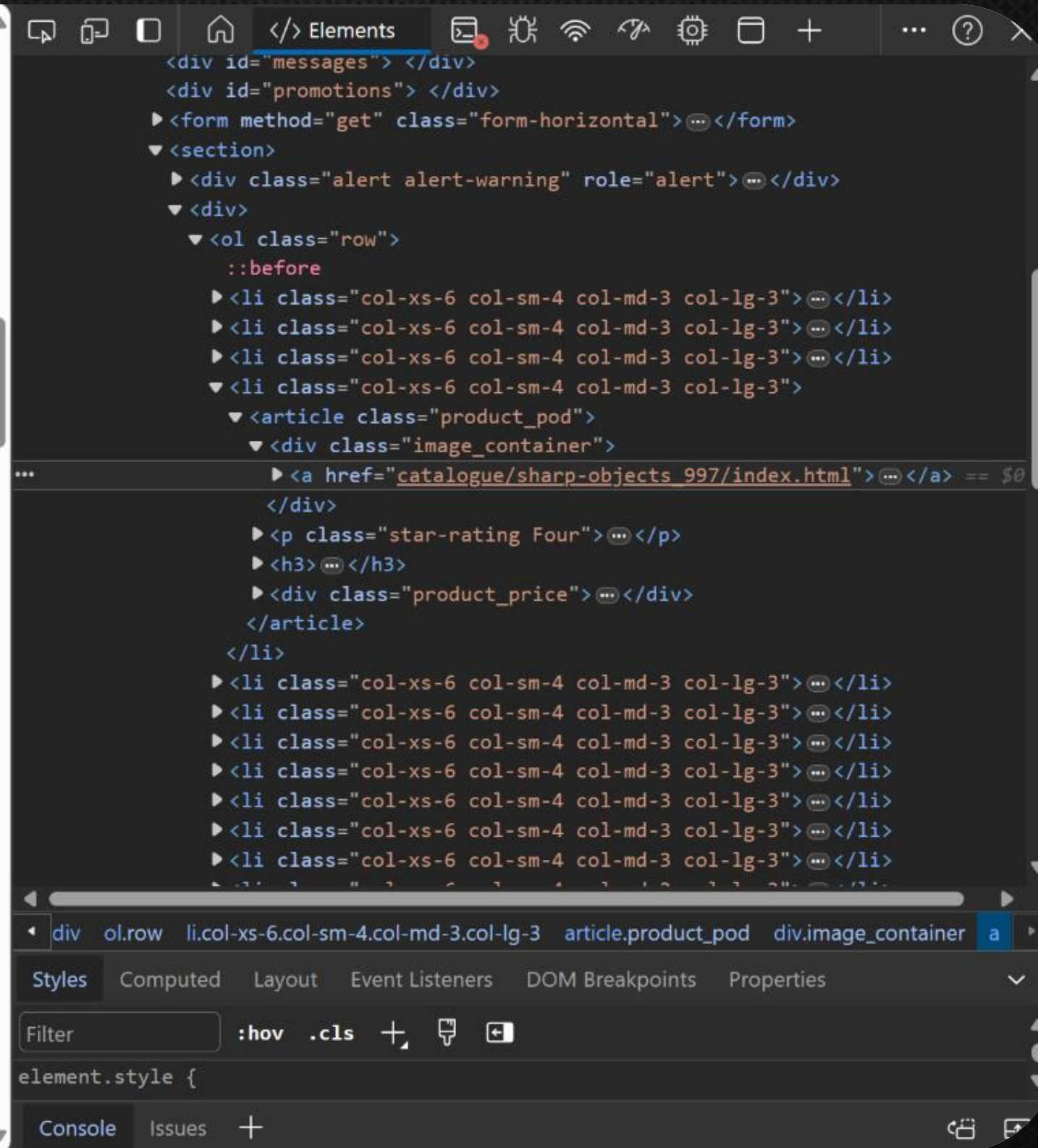
Elements

div   ol.row   li.col-xs-6.col-sm-4.col-md-3.col-lg-3   article.product_pod   div.image_container   a

Styles   Computed   Layout   Event Listeners   DOM Breakpoints   Properties

Filter   :hov   .cls

element.style {
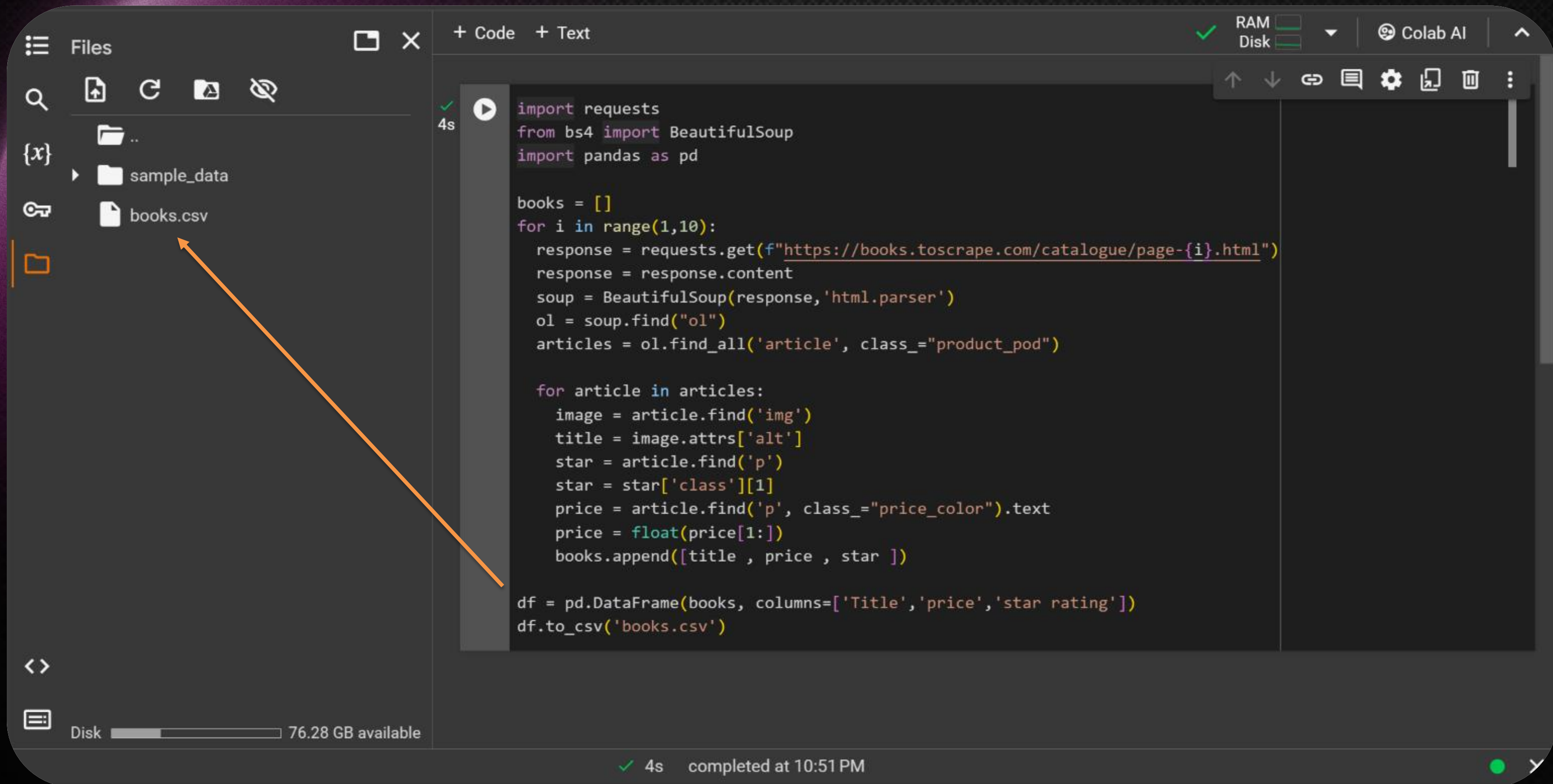
Console   Issues

This is the main web page from which we extracted the data

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

books = []
for i in range(1,10):
  response = requests.get(f"https://books.toscrape.com/catalogue/page-{i}.html")
  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']
    star = article.find('p')
    star = star['class'][1]
    price = article.find('p', class_="price_color").text
    price = float(price[1:])
    books.append([title , price , star ])

df = pd.DataFrame(books, columns=['Title','price','star rating'])
df.to_csv('books.csv')
```

✓ Connected to Python 3 Google Compute Engine backend

The code

Program output

Files

- ..
- ▶ sample_data
- books.csv

Disk ▬▬▬▬▬▬▬ 76.28 GB available

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

books = []
for i in range(1,10):
  response = requests.get(f"https://books.toscrape.com/catalogue/page-{i}.html")
  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']
    star = article.find('p')
    star = star['class'][1]
    price = article.find('p', class_="price_color").text
    price = float(price[1:])
    books.append([title , price , star ])

df = pd.DataFrame(books, columns=['Title','price','star rating'])
df.to_csv('books.csv')
```

✓ 4s completed at 10:51 PM

Program output

sample_data

books.csv

```
import req
from bs4 i
import pan

books = []
for i in r
    respons
    respons
    soup = B
    ol = sou
    articles

    for arti
        image
        title
        star =
        star =
        price
        price
        books.

df = pd.Da
df.to_csv(
```

4s

books.csv

	Title	price	star rating
0	A Light in the Attic	51.77	Three
1	Tipping the Velvet	53.74	One
2	Soumission	50.1	One
3	Sharp Objects	47.82	Four
4	Sapiens: A Brief History of Humankind	54.23	Five
5	The Requiem Red	22.65	One
6	The Dirty Little Secrets of Getting Your Dream Job	33.34	Four
7	The Coming Woman: A Novel Based on the Life of the Infamous Feminist, Victoria Woodhull	17.93	Three
8	The Boys in the Boat: Nine Americans and Their Epic Quest for Gold at the 1936 Berlin Olympics	22.6	Four
9	The Black Maria	52.15	One
10	Starving Hearts (Triangular Trade Trilogy, #1)	13.99	Two
11	Shakespeare's Sonnets	20.66	Four
12	Set Me Free	17.46	Five
13	Scott Pilgrim's Precious Little Life (Scott Pilgrim #1)	52.29	Five
14	Rip it Up and Start Again	35.02	Five
15	Our Band Could Be Your Life: Scenes from the American Indie Underground, 1981-1991	57.25	Three
16	Olio	23.88	One
17	Mesaerion: The Best Science Fiction Stories 1800-1849	37.59	One
18	Libertarianism for Beginners	51.33	Two
19	It's Only the Himalayas	45.17	Two
20	In Her Wake	12.84	One
21	How Music Works	37.32	Two
22	Foolproof Preserving: A Guide to Small Batch Jams, Jellies, Pickles, Condiments, and More: A Foolproof Guide to Making Small Batch Jams, Jellies, Pickles, Condiments, and More	30.52	Three
23	Chase Me (Paris Nights #2)	25.27	Five
24	Black Dust	34.53	Five

Show 25 per page   1 2 3 4 5 6 7 8

Disk ▬▬▬▬▬▬▬▬▬ 81.65 GB available

+ Code   + Text   All changes saved   RAM Disk   Colab AI

Paste | Cut | Copy | Format Painter
Clipboard

Bahnschrift 11 | B I U
Font

Wrap Text | Merge & Center
Alignment

General | $ % , .00 .0
Number

Normal | Bad | Good
Neutral | Calculation | Check Cell
Conditional Formatting | Format as Table
Styles

Insert | Delete | Format
Cells

AutoSum | Fill | Clear
Sort & Filter | Find & Select
Editing

B2 | A Light in the Attic

	A	B	C	D	E	F
1		Title	price	star rating		
2	0	A Light in the Attic	51.77	Three		
3	1	Tipping the Velvet	53.74	One		
4	2	Soumission	50.1	One		
5	3	Sharp Objects	47.82	Four		
6	4	Sapiens: A Brief History of Humankind	54.23	Five		
7	5	The Requiem Red	22.65	One		
8	6	The Dirty Little Secrets of Getting Your Dream Job	33.34	Four		
9	7	The Coming Woman: A Novel Based on the Life of the Infamous Feminist, Victoria Woodhull	17.93	Three		
10	8	The Boys in the Boat: Nine Americans and Their Epic Quest for Gold at the 1936 Berlin Olympics	22.6	Four		
11	9	The Black Maria	52.15	One		
12	10	Starving Hearts (Triangular Trade Trilogy, #1)	13.99	Two		
13	11	Shakespeare's Sonnets	20.66	Four		
14	12	Set Me Free	17.46	Five		
15	13	Scott Pilgrim's Precious Little Life (Scott Pilgrim #1)	52.29	Five		
16	14	Rip it Up and Start Again	35.02	Five		
17	15	Our Band Could Be Your Life: Scenes from the American Indie Underground, 1981-1991	57.25	Three		
18	16	Olio	23.88	One		
19	17	Mesaerion: The Best Science Fiction Stories 1800-1849	37.59	One		
20	18	Libertarianism for Beginners	51.33	Two		
21	19	It's Only the Himalayas	45.17	Two		
22	20	In Her Wake	12.84	One		
23	21	How Music Works	37.32	Two		
24	22	Foolproof Preserving: A Guide to Small Batch Jams, Jellies, Pickles, Condiments, and More: A Foolproof Guide to Making Small Batch Jams, Jellies, Pickles, Condimen	30.52	Three		
25	23	Chase Me (Paris Nights #2)	25.27	Five		
26	24	Black Dust	34.53	Five		
27	25	Birdsong: A Story in Pictures	54.64	Three		
28	26	America's Cradle of Quarterbacks: Western Pennsylvania's Football Factory from Johnny Unitas to Joe Montana	22.5	Three		
29	27	Aladdin and His Wonderful Lamp	53.13	Three		
30	28	Worlds Elsewhere: Journeys Around Shakespeare's Globe	40.3	Five		
31	29	Wall and Piece	44.18	Four		
32	30	The Four Agreements: A Practical Guide to Personal Freedom	17.66	Five		

بيانات المشروع

After extracting the data and placing it in a csv file, it is transferred to Excel to benefit from it more easily

Thanks

UOITC