Amazon Audio Book Web Scraping using BeautifulSoup

In [4]:
```
#Importing Request and Bs4 Module
import requests
from bs4 import BeautifulSoup as bs
```

```
C:\Users\PUJA\anaconda3\lib\site-packages\requests\__init__.py:91: RequestsDependencyWarning: urllib3 (1.26.12) or char
det (3.0.4) doesn't match a supported version!
  RequestsDependencyWarning)
```

website which we want to scrap stored is variable Link

In [5]:
```
Link="https://www.audible.in/adblbestsellers?ref=a_ep_free_t1_navTop_pl1cg0c1r0&pf_rd_p=45682ad3-a4ca-473f-b9dc-aedd02f1e
```

Get th HTML

In [6]:
```
page=requests.get(Link)  #Get th HTML
page.content
```

...

Parse the html

In [7]:
```
soup=bs(page.content,"html.parser")   #Parse the html
soup
```

```
3s linear both;animation:hide .5s .3s linear both}.flip-clock-wrapper ul.play li.flip-clock-before .down .shadow{bac
kground:-moz-linear-gradient(top,#000 0,rgba(0,0,0,.1) 100%);background:-webkit-gradient(linear,left top,left botto
m,color-stop(0,#000),color-stop(100%,rgba(0,0,0,.1)));background:linear,top,#000 0,rgba(0,0,0,.1) 100%;background:-o
-linear-gradient(top,#000 0,rgba(0,0,0,.1) 100%);background:-ms-linear-gradient(top,#000 0,rgba(0,0,0,.1) 100%);back
ground:linear,to bottom,#000 0,rgba(0,0,0,.1) 100%;-webkit-animation:show .5s linear both;-moz-animation:show .5s li
near both;animation:show .5s linear both}.flip-clock-wrapper ul.play li.flip-clock-active .down .shadow{background:-
moz-linear-gradient(top,#000 0,rgba(0,0,0,.1) 100%);background:-webkit-gradient(linear,left top,left bottom,color-st
op(0,#000),color-stop(100%,rgba(0,0,0,.1)));background:linear,top,#000 0,rgba(0,0,0,.1) 100%;background:-o-linear-gr
adient(top,#000 0,rgba(0,0,0,.1) 100%);background:-ms-linear-gradient(top,#000 0,rgba(0,0,0,.1) 100%);background:lin
ear,to bottom,#000 0,rgba(0,0,0,.1) 100%;-webkit-animation:hide .5s .3s linear both;-moz-animation:hide .5s .3s line
ar both;animation:hide .5s .2s linear both}@-webkit-keyframes show{0%{opacity:0}100%{opacity:1}}@-moz-keyframes show
{0%{opacity:0}100%{opacity:1}}@-o-keyframes show{0%{opacity:0}100%{opacity:1}}@keyframes show{0%{opacity:0}100%{opac
ity:1}}@-webkit-keyframes hide{0%{opacity:1}100%{opacity:0}}@-moz-keyframes hide{0%{opacity:1}100%{opacity:0}}@-o-ke
yframes hide{0%{opacity:1}100%{opacity:0}}@keyframes hide{0%{opacity:1}100%{opacity:0}}.easyExchangeContainer{backgr
ound-position-x:center!important}.easyExchangeAsinImage{width:192px}.greetingHeaders .bc-heading{font-weight:400}.gr
eetingSubHeaders .bc-text{font-weight:300}.greetingHeaders{text-align:center}.greetingSubHeaders{text-align:center}.
dlp-header{background:0 0;position:absolute;left:0;right:0;z-index:10}.dlp-modal-bg{position:fixed;height:100%;widt
h:100%;top:0;left:0;background:rgba(0,0,0,.7);z-index:20;display:none}.dlp-modal{position:fixed;z-index:21;top:100p
x;width:830px;height:0;left:0;right:0;margin:auto;display:none}.dlp-modal .close{position:absolute;right:20px;top:20
px;cursor:pointer}.a-marketplace:hover{opacity:.5;filter:alpha(opacity=50)}.anon-mbrshp-button-container{width:320p
```

In [8]:
```
print(soup.prettify())
```

...

In [9]:
```
list(soup.children)
```

...

title of the page

In [10]:
```
title=soup.title
title
```

Out[10]:
```
<title>Our Best Audiobooks: Most Popular &amp;amp; Best Sellers| Audible.in</title>
```

In [11]:
```
print(type(soup))
```

```
<class 'bs4.BeautifulSoup'>
```

In [12]:
```
print(type(title))
```

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

Title of the Page

In [13]:
```
print(title.string)
```

```
Our Best Audiobooks: Most Popular & Best Sellers| Audible.in
```

created Function for cleaning data

In [14]:
```
def Rnewline(lst):
    new=[]
    for i in lst:
        new.append(i.strip().replace("\n",""))
    return new
```

Get all the Audio book Name

In [15]:
```
name=soup.find_all("h3",class_="bc-heading bc-color-link bc-pub-break-word bc-size-medium")
name
```

...

In [16]:
```
name1=[]
for i in range(0,len(name)):
    name1.append((name[i].get_text()).replace(" ",""))

AU_name=Rnewline(name1)
AU_name
```

Out[16]:
```
['1.AtomicHabits',
 '2.ThePsychologyofMoney',
 '3.HowtoTalktoAnyone',
 '4.RichDadPoorDad',
 '5.TheSubtleArtofNotGivingaF*ck',
 '6.Ikigai',
 '7.TheCouragetoBeDisliked',
 "8.Life'sAmazingSecrets",
 '9.Sapiens',
 '10.Hyperfocus',
 '11.TheAlmanackofNavalRavikant',
 '12.WarofLanka',
 '13.Meditations',
 "14.Can'tHurtMe",
 '15.48LawsofPower',
 '16.NeverSplittheDifference',
 '17.Thinking,FastandSlow',
 '18.MasterYourEmotions:APracticalGuidetoOvercomeNegativityandBetterManageYourFeelings',
 '19.TheIntelligentInvestorRevEd.',
 '20.HowtoWinFriendsandInfluencePeople']
```

Get the writer of the book

In [17]:
```
wrt=soup.find_all("li",class_="bc-list-item authorLabel")
wrt
```

Out[17]:
```
[<li class="bc-list-item authorLabel">
 <span class="bc-text bc-size-small bc-color-secondary">
                                    Written by:
                                    <a class="bc-link bc-color-link" href="/author/James-Clear/B07DJTJC3X" tabindex
="0">James Clear</a>
 </span>
 </li>,
 <li class="bc-list-item authorLabel">
 <span class="bc-text bc-size-small bc-color-secondary">
                                    Written by:
                                    <a class="bc-link bc-color-link" href="/author/Morgan-Housel/B084ZNV1LR" tabind
ex="0">Morgan Housel</a>
 </span>
 </li>,
 <li class="bc-list-item authorLabel">
 <span class="bc-text bc-size-small bc-color-secondary">
                                    Written by:
                                    <a class="bc-link bc-color-link" href="/author/Leil-Lowndes/B000APOPHO" tabinde
x="0">Leil Lowndes</a>
```

In [37]:
```
wrt1=[]
for i in range(0,len(wrt)):
    wrt1.append((wrt[i].get_text()).replace("Written by:\n", ""))

AU_wrt=Rnewline(wrt1)
AU_wrt
```

Out[37]:
```
['James Clear',
 'Morgan Housel',
 'Leil Lowndes',
 'Robert T. Kiyosaki',
 'Mark Manson',
 'Héctor García, Francesc Miralles',
 'Fumitake Koga, Ichiro Kishimi',
 'Gaur Gopal Das',
 'Yuval Noah Harari',
 'Chris Bailey',
 'Eric Jorgenson, Tim Ferriss',
 'Amish Tripathi',
 'Audible Sleep',
 'David Goggins',
 'Robert Greene',
 'Chris Voss, Tahl Raz',
 'Daniel Kahneman',
 'Thibaut Meurisse',
 'Benjamin Graham',
 'Dale Carnegie']
```

Get the Audio book length

In [19]:
```
time=soup.find_all("li",class_="bc-list-item runtimeLabel")
time
```

...

In [20]:
```
time1=[]
for i in range(0,len(time)):
    time1.append((time[i].get_text()).replace("Length:",""))
time1
```

Out[20]:
```
['\n 5 hrs and 35 mins\n',
 '\n 5 hrs and 48 mins\n',
 '\n 8 hrs and 59 mins\n',
 '\n 6 hrs and 9 mins\n',
 '\n 5 hrs and 17 mins\n',
 '\n 3 hrs and 23 mins\n',
 '\n 6 hrs and 50 mins\n',
 '\n 6 hrs and 25 mins\n',
 '\n 15 hrs and 18 mins\n',
 '\n 6 hrs and 39 mins\n',
 '\n 4 hrs and 53 mins\n',
 '\n 16 hrs and 20 mins\n',
 '\n Not Yet Known\n',
 '\n 13 hrs and 37 mins\n',
 '\n 23 hrs and 6 mins\n',
 '\n 8 hrs and 7 mins\n',
 '\n 20 hrs and 1 min\n',
 '\n 3 hrs and 54 mins\n',
 '\n 17 hrs and 48 mins\n',
 '\n 8 hrs and 5 mins\n']
```

In [21]:
```
AU_time= Rnewline(time1)
```

In [22]:
```
AU_time
```

Out[22]:
```
['5 hrs and 35 mins',
 '5 hrs and 48 mins',
 '8 hrs and 59 mins',
 '6 hrs and 9 mins',
 '5 hrs and 17 mins',
 '3 hrs and 23 mins',
 '6 hrs and 50 mins',
 '6 hrs and 25 mins',
 '15 hrs and 18 mins',
 '6 hrs and 39 mins',
 '4 hrs and 53 mins',
 '16 hrs and 20 mins',
 'Not Yet Known',
 '13 hrs and 37 mins',
 '23 hrs and 6 mins',
 '8 hrs and 7 mins',
 '20 hrs and 1 min',
 '3 hrs and 54 mins',
 '17 hrs and 48 mins',
 '8 hrs and 5 mins']
```

Get the Audio book Language

In [23]:
```
lan=soup.find_all("li",class_="bc-list-item languageLabel")
lan
```

...

In [24]:
```
lan1=[]
for i in range(0,len(lan)):
    lan1.append((lan[i].get_text()).replace("Language:",""))

len(lan1)
```

Out[24]:
```
20
```

In [25]:
```
AU_lan= Rnewline(lan1)
AU_lan
```

Out[25]:
```
['English',
 'English',
 'English',
 'English',
 'English',
 'English',
 'English',
 'English',
 'English',
 'English',
 'English',
 'English',
 'English',
 'English',
 'English',
 'English',
 'English',
 'English',
 'English',
 'English',
 'English']
```

Get the Audio book Release date

In [26]:
```
rdate=soup.find_all("li",class_="bc-list-item releaseDateLabel")
rdate
```

...

In [27]:
```
rdate1=[]
for i in range(0, len(rdate)):
    rdate1.append((rdate[i].get_text()).replace("Release Date:",""))
```

In [28]:
```
AU_Rdate=Rnewline(rdate1)
AU_Rdate
```

Out[28]:
```
['18-10-18',
 '08-09-20',
 '01-09-15',
 '15-06-12',
 '13-09-16',
 '18-09-17',
 '22-02-18',
 '10-12-19',
 '30-04-15',
 '06-09-18',
 '01-03-21',
 '03-10-22',
 '01-06-20',
 '28-11-18',
 '01-05-15',
 '20-06-19',
 '23-12-11',
 '01-10-19',
 '07-07-15',
 '08-11-18']
```

Get the Audio book Rating

In [29]:
```
rate=soup.find_all("li",class_="bc-list-item ratingsLabel")
rate
```

...

In [30]:
```
rate1=[]
for i in range(0,len(rate)):
    rate1.append(rate[i].get_text())

AU_rating=Rnewline(rate1)
AU_rating
```

Out[30]:
```
['5 out of 5 stars15,003 ratings',
 '4.5 out of 5 stars7,666 ratings',
 '4.5 out of 5 stars795 ratings',
 '4.5 out of 5 stars6,026 ratings',
 '4.5 out of 5 stars4,286 ratings',
 '4.5 out of 5 stars7,763 ratings',
 '4.5 out of 5 stars862 ratings',
 '4.5 out of 5 stars4,927 ratings',
 '5 out of 5 stars6,037 ratings',
 '4.5 out of 5 stars1,158 ratings',
 '4.5 out of 5 stars1,343 ratings',
 '4.5 out of 5 stars325 ratings',
 '4 out of 5 stars128 ratings',
 '5 out of 5 stars2,835 ratings',
 '4.5 out of 5 stars373 ratings',
 '4.5 out of 5 stars1,099 ratings',
 '4.5 out of 5 stars775 ratings',
 '4.5 out of 5 stars203 ratings',
 '4.5 out of 5 stars365 ratings',
 '4.5 out of 5 stars2,648 ratings']
```

Get the Price of the book

In [31]:
```
price=soup.find_all("span",class_="bc-text bc-size-base bc-color-base")
price
```

...

In [32]:
```
price1=[]
for i in range(0,len(price)):
    price1.append((price[i].get_text()).replace("Sale price:","").replace("Regular price:","").replace("or 1 credit","")

r=Rnewline(price1)
r
#len(r)
```

...

In [33]:
```
AU_price=[]
for i in range(1,60,3):
    AU_price.append(r[i])

AU_price
```

Out[33]:
```
['₹820.00',
 '₹668.00',
 '₹844.00',
 '₹844.00',
 '₹1,238.00',
 '₹615.00',
 '₹1,063.00',
 '₹1,005.00',
 '₹957.00',
 '₹323.00',
 '₹586.00',
 '₹1,003.00',
 '₹836.00',
 '₹836.00',
 '₹820.00',
 '₹820.00',
 '₹957.00',
 '₹501.00',
 '₹2,194.00',
 '₹155.00']
```

Importing into Panda for Frame Data

In [34]:
```
import pandas as pd
```

In [35]:
```
df=pd.DataFrame()
df["AUDIO_BOOK_NAMES"]=AU_name
df["writer"]=AU_wrt
df["Release_date"]=AU_Rdate
df["Language"]=AU_lan
df["Audio_Lenghth"]=AU_time
df["Rating"]=AU_rating
df["Audio_Lenghth"]=AU_time
df["Price"]=AU_price
```

In [36]:
```
df
```

Out[36]:

	AUDIO_BOOK_NAMES	writer	Release_date	Language	Audio_Lenghth	Rating	Price
0	1.AtomicHabits	James Clear	18-10-18	English	5 hrs and 35 mins	5 out of 5 stars15,003 ratings	₹820.00
1	2.ThePsychologyofMoney	Morgan Housel	08-09-20	English	5 hrs and 48 mins	4.5 out of 5 stars7,666 ratings	₹668.00
2	3.HowtoTalktoAnyone	Leil Lowndes	01-09-15	English	8 hrs and 59 mins	4.5 out of 5 stars795 ratings	₹844.00
3	4.RichDadPoorDad	Robert T. Kiyosaki	15-06-12	English	6 hrs and 9 mins	4.5 out of 5 stars6,026 ratings	₹844.00
4	5.TheSubtleArtofNotGivingaF*ck	Mark Manson	13-09-16	English	5 hrs and 17 mins	4.5 out of 5 stars4,286 ratings	₹1,238.00
5	6.Ikigai	Héctor García, Francesc Miralles	18-09-17	English	3 hrs and 23 mins	4.5 out of 5 stars7,763 ratings	₹615.00
6	7.TheCouragetoBeDisliked	Fumitake Koga, Ichiro Kishimi	22-02-18	English	6 hrs and 50 mins	4.5 out of 5 stars862 ratings	₹1,063.00
7	8.Life'sAmazingSecrets	Gaur Gopal Das	10-12-19	English	6 hrs and 25 mins	4.5 out of 5 stars4,927 ratings	₹1,005.00
8	9.Sapiens	Yuval Noah Harari	30-04-15	English	15 hrs and 18 mins	5 out of 5 stars6,037 ratings	₹957.00
9	10.Hyperfocus	Chris Bailey	06-09-18	English	6 hrs and 39 mins	4.5 out of 5 stars1,158 ratings	₹323.00
10	11.TheAlmanackofNavalRavikant	Eric Jorgenson, Tim Ferriss	01-03-21	English	4 hrs and 53 mins	4.5 out of 5 stars1,343 ratings	₹586.00
11	12.WarofLanka	Amish Tripathi	03-10-22	English	16 hrs and 20 mins	4.5 out of 5 stars325 ratings	₹1,003.00
12	13.Meditations	Audible Sleep	01-06-20	English	Not Yet Known	4 out of 5 stars128 ratings	₹836.00
13	14.Can'tHurtMe	David Goggins	28-11-18	English	13 hrs and 37 mins	5 out of 5 stars2,835 ratings	₹836.00
14	15.48LawsofPower	Robert Greene	01-05-15	English	23 hrs and 6 mins	4.5 out of 5 stars373 ratings	₹820.00

	AUDIO_BOOK_NAMES	writer	Release_date	Language	Audio_Lenghth	Rating	Price
15	16.NeverSplittheDifference	Chris Voss, Tahl Raz	20-06-19	English	8 hrs and 7 mins	4.5 out of 5 stars1,099 ratings	₹820.00
16	17.Thinking,FastandSlow	Daniel Kahneman	23-12-11	English	20 hrs and 1 min	4.5 out of 5 stars775 ratings	₹957.00
17	18.MasterYourEmotions:APracticalGuidetoOvercom...	Thibaut Meurisse	01-10-19	English	3 hrs and 54 mins	4.5 out of 5 stars203 ratings	₹501.00
18	19.TheIntelligentInvestorRevEd.	Benjamin Graham	07-07-15	English	17 hrs and 48 mins	4.5 out of 5 stars365 ratings	₹2,194.00
19	20.HowtoWinFriendsandInfluencePeople	Dale Carnegie	08-11-18	English	8 hrs and 5 mins	4.5 out of 5 stars2,648 ratings	₹155.00

Creating CSV file to store the data

In [35]:
```
df.to_csv("Best Amazon audio book ",index=0)
```

In [ ]:
```

```