WEB SCRAPPING ASSIGNMENT 4

Question 1: Scrape the details of most viewed videos on YouTube from Wikipedia. Url = https://en.wikipedia.org/wiki/List_of_most-viewed_YouTube_videos (https://en.wikipedia.org/wiki/List_of_most-viewed_YouTube_videos) You need to find following details:

In [1]:
```
!pip install selenium
```

```
Requirement already satisfied: selenium in c:\users\dell\anaconda3\lib\site-packages (4.15.2)
Requirement already satisfied: trio~=0.17 in c:\users\dell\anaconda3\lib\site-packages (from selenium) (0.2
3.1)
Requirement already satisfied: urllib3[socks]<3,>=1.26 in c:\users\dell\anaconda3\lib\site-packages (from se
lenium) (1.26.14)
Requirement already satisfied: certifi>=2021.10.8 in c:\users\dell\anaconda3\lib\site-packages (from seleniu
m) (2022.12.7)
Requirement already satisfied: trio-websocket~=0.9 in c:\users\dell\anaconda3\lib\site-packages (from seleni
um) (0.11.1)
Requirement already satisfied: sniffio>=1.3.0 in c:\users\dell\anaconda3\lib\site-packages (from trio~=0.17-
>selenium) (1.3.0)
Requirement already satisfied: outcome in c:\users\dell\anaconda3\lib\site-packages (from trio~=0.17->seleni
um) (1.3.0.post0)
Requirement already satisfied: exceptiongroup>=1.0.0rc9 in c:\users\dell\anaconda3\lib\site-packages (from t
rio~=0.17->selenium) (1.1.3)
Requirement already satisfied: attrs>=20.1.0 in c:\users\dell\anaconda3\lib\site-packages (from trio~=0.17->
selenium) (22.1.0)
Requirement already satisfied: sortedcontainers in c:\users\dell\anaconda3\lib\site-packages (from trio~=0.1
7->selenium) (2.4.0)
Requirement already satisfied: cffi>=1.14 in c:\users\dell\anaconda3\lib\site-packages (from trio~=0.17->sel
enium) (1.15.1)
Requirement already satisfied: idna in c:\users\dell\anaconda3\lib\site-packages (from trio~=0.17->selenium)
(3.4)
Requirement already satisfied: wsproto>=0.14 in c:\users\dell\anaconda3\lib\site-packages (from trio-websock
et~=0.9->selenium) (1.2.0)
Requirement already satisfied: PySocks!=1.5.7,<2.0,>=1.5.6 in c:\users\dell\anaconda3\lib\site-packages (fro
m urllib3[socks]<3,>=1.26->selenium) (1.7.1)
Requirement already satisfied: pycparser in c:\users\dell\anaconda3\lib\site-packages (from cffi>=1.14->trio
~=0.17->selenium) (2.21)
Requirement already satisfied: h11<1,>=0.9.0 in c:\users\dell\anaconda3\lib\site-packages (from wsproto>=0.1
4->trio-websocket~=0.9->selenium) (0.14.0)
```

In [2]:
```
import selenium
from selenium import webdriver
import pandas as pd
from selenium.webdriver.common.by import By
import warnings
warnings.filterwarnings("ignore")
import time
from selenium.common.exceptions import NoSuchElementException
from selenium.common.exceptions import ElementClickInterceptedException
```

In [3]:
```
driver = webdriver.Chrome()
```

In [4]:
```
driver.get("https://en.wikipedia.org/wiki/List_of_most-viewed_YouTube_videos")
```

In [5]:
```
rank =[]
name =[]
artist =[]
upload = []
views =[]

Rank  = driver.find_elements(By.XPATH,'//table[@class="wikitable sortable jquery-tablesorter"]//tbody//tr//td
for i in Rank:
    rank.append(i.text)
Name  = driver.find_elements(By.XPATH,'//table[@class="wikitable sortable jquery-tablesorter"]//tbody//tr//td
for i in Name:
    name.append(i.text)
Artists = driver.find_elements(By.XPATH,'//table[@class="wikitable sortable jquery-tablesorter"]//tbody//tr//
for i in Artists:
    artist.append(i.text)
Upload_date = driver.find_elements(By.XPATH,'//table[@class="wikitable sortable jquery-tablesorter"]//tbody//
for i in Upload_date:
    upload.append(i.text)
Views = driver.find_elements(By.XPATH,'//table[@class="wikitable sortable jquery-tablesorter"]//tbody//tr//td
for i in Views:
    views.append(i.text)
```

In [6]:
```
df = pd.DataFrame({'RANK':rank,'Video Name':name,'Artist Name':artist,'Upload date':upload,'Views in Billion'
df.head(30)
```

Out[6]:

	RANK	Video Name	Artist Name	Upload date	Views in Billion
0	1.	"Baby Shark Dance"[6]	Pinkfong Baby Shark - Kids' Songs & Stories	June 17, 2016	13.65
1	2.	"Despacito"[9]	Luis Fonsi	January 12, 2017	8.32
2	3.	"Johny Johny Yes Papa"[17]	LooLoo Kids - Nursery Rhymes and Children's Songs	October 8, 2016	6.84
3	4.	"Bath Song"[18]	Cocomelon - Nursery Rhymes	May 2, 2018	6.50
4	5.	"Shape of You"[19]	Ed Sheeran	January 30, 2017	6.14
5	6.	"See You Again"[22]	Wiz Khalifa	April 6, 2015	6.09
6	7.	"Wheels on the Bus"[27]	Cocomelon - Nursery Rhymes	May 24, 2018	5.71
7	8.	"Phonics Song with Two Words"[28]	ChuChu TV Nursery Rhymes & Kids Songs	March 6, 2014	5.57
8	9.	"Uptown Funk"[29]	Mark Ronson	November 19, 2014	5.09
9	10.	"Learning Colors – Colorful Eggs on a Farm"[30]	Miroshka TV	February 27, 2018	5.01
10	11.	"Gangnam Style"[31]	officialpsy	July 15, 2012	4.96
11	12.	"Masha and the Bear – Recipe for Disaster"[36]	Get Movies	January 31, 2012	4.57
12	13.	"Dame Tu Cosita"[37]	Ultra Records	April 5, 2018	4.48
13	14.	"Axel F"[38]	Crazy Frog	June 16, 2009	4.16
14	15.	"Sugar"[39]	Maroon 5	January 14, 2015	3.97
15	16.	"Counting Stars"[40]	OneRepublic	May 31, 2013	3.92
16	17.	"Roar"[41]	Katy Perry	September 5, 2013	3.91
17	18.	"Baa Baa Black Sheep"[42]	Cocomelon - Nursery Rhymes	June 25, 2018	3.84
18	19.	"Waka Waka (This Time for Africa)"[43]	Shakira	June 4, 2010	3.78
19	20.	"Lakdi Ki Kathi"[44]	Jingle Toons	June 14, 2018	3.76
20	21.	"Sorry"[45]	Justin Bieber	October 22, 2015	3.74
21	22.	"Thinking Out Loud"[46]	Ed Sheeran	October 7, 2014	3.69
22	23.	"Humpty the train on a fruits ride"[47]	Kiddiestv Hindi - Nursery Rhymes & Kids Songs	January 26, 2018	3.63
23	24.	"Dark Horse"[48]	Katy Perry	February 20, 2014	3.63

	RANK	Video Name	Artist Name	Upload date	Views in Billion
24	25.	"Perfect"[49]	Ed Sheeran	November 9, 2017	3.60
25	26.	"Let Her Go"[50]	Passenger	July 25, 2012	3.56
26	27.	"Faded"[51]	Alan Walker	December 3, 2015	3.55
27	28.	"Shree Hanuman Chalisa"[52]	T-Series Bhakti Sagar	May 10, 2011	3.54
28	29.	"Girls Like You"[53]	Maroon 5	May 31, 2018	3.52
29	30.	"Lean On"[54]	Major Lazer Official	March 22, 2015	3.50

In [7]:
```
driver.close()
```

Question 2: Scrape the details team India's international fixtures from bcci.tv. Url = https://www.bcci.tv/ (https://www.bcci.tv/). You need to find following details:

In [8]:
```
driver = webdriver.Chrome()
driver.get("https://www.bcci.tv/")
fixure = driver.find_element(By.XPATH,'//div[@class="imw-tabs international-tabs"]//a[2]')
fixure.click()
```

In [9]:
```
series =[]
place=[]
date = []
time =[]

Series = driver.find_elements(By.XPATH,'//div[@class="match-card match-card-fw match-card-up ng-scope"]//div[
for i in Series:
    series.append(i.text)
Place = driver.find_elements(By.XPATH,'//div[@class="match-place ng-scope"]')
for i in Place:
    place.append(i.text)
Date  = driver.find_elements(By.XPATH,'//div[@class="match-dates ng-binding"]')
for i in Date:
    date.append(i.text)
Time = driver.find_elements(By.XPATH,'//div[@class="match-time no-margin ng-binding"]')
for i in Time:
    time.append(i.text)
```

In [10]:
```
series
```

Out[10]:
```
["Women's T20 Match",
 '4th T20I',
 "Women's T20 Match",
 '5th T20I',
 '1st T20I',
 '2nd T20I',
 '3rd T20I',
 '1st Test',
 '1st Test']
```

In [11]:
```
print(len(series),len(place),len(date),len(time))
```

```
9 9 9 9
```

In [12]:
```
df = pd.DataFrame({'Series':series,'Place':place})
df1 = pd.DataFrame({"Date":date,'Time':time})
df2 = pd.concat([df,df1],axis=1)
df2
```

Out[12]:

	Series	Place	Date	Time
0	Women's T20 Match	Wankhede Stadium, Mumbai	1 DECEMBER, 2023	1:30 PM IST
1	4th T20I	Shaheed Veer Narayan Singh International Crick...	1 DECEMBER, 2023	7:00 PM IST
2	Women's T20 Match	Wankhede Stadium, Mumbai	3 DECEMBER, 2023	1:30 PM IST
3	5th T20I	M Chinnaswamy Stadium, Bengaluru	3 DECEMBER, 2023	7:00 PM IST
4	1st T20I	Wankhede Stadium, Mumbai	6 DECEMBER, 2023	7:00 PM IST
5	2nd T20I	Wankhede Stadium, Mumbai	9 DECEMBER, 2023	7:00 PM IST
6	3rd T20I	Wankhede Stadium, Mumbai	10 DECEMBER, 2023	7:00 PM IST
7	1st Test	DY Patil Stadium, NAVI MUMBAI	14 DECEMBER, 2023	9:30 AM IST
8	1st Test	Wankhede Stadium, Mumbai	21 DECEMBER, 2023	9:30 AM IST

In [13]:
```
driver.close()
```

Question 3:Scrape the details of State-wise GDP of India from statisticstime.com. Url = http://statisticstimes.com/ (http://statisticstimes.com/)

In [14]:
```
driver = webdriver.Chrome()
driver.get("http://statisticstimes.com/")
```

In [15]:
```
economy = driver.find_element(By.XPATH,'//div[@class="navbar"]//div[2]')
economy.click()
india = driver.find_element(By.XPATH,'//div[@class="navbar"]//div[2]//div//a[3]')
india.click()
```

In [16]:
```
try:
    gdp = driver.find_element(By.XPATH,'//div[@style="float:left;width:1150px;height:800px;background-color:
    gdp.click()

except NoSuchElementException as e:

    print("Exception Raised: ", e)
```

In [17]:
```
rank =[]
state = []
gsdp1 =[]
gsdp2 =[]
share = []
gdp =[]

Rank  = driver.find_elements(By.XPATH,'//div[@class="fwidth"][3]//div//table//tbody//tr//td[1]')
for i in Rank:
    rank.append(i.text)
Sname  = driver.find_elements(By.XPATH,'//div[@class="fwidth"][3]//div//table//tbody//tr//td[2]')
for i in Sname:
    state.append(i.text)
GDP20 = driver.find_elements(By.XPATH,'//div[@class="fwidth"][3]//div//table//tbody//tr//td[3]')
for i in GDP20:
    gsdp1.append(i.text)
GDP19 = driver.find_elements(By.XPATH,'//div[@class="fwidth"][3]//div//table//tbody//tr//td[4]')
for i in GDP19:
    gsdp2.append(i.text)
Share = driver.find_elements(By.XPATH,'//div[@class="fwidth"][3]//div//table//tbody//tr//td[5]')
for i in Share:
    share.append(i.text)
GDP = driver.find_elements(By.XPATH,'//div[@class="fwidth"][3]//div//table//tbody//tr//td[6]')
for i in GDP:
    gdp.append(i.text)
```

In [18]:
```
df= pd.DataFrame({'Rank':rank,'State':state,'GSDP(18-19)':gsdp2,'GSDP(19-20)':gsdp1,'Share(18-19)':share,'GDP
df
```

Out[18]:

	Rank	State	GSDP(18-19)	GSDP(19-20)	Share(18-19)	GDP($Billion)
0	1	Maharashtra	2,632,792	-	13.94%	399.921
1	2	Tamil Nadu	1,630,208	1,845,853	8.63%	247.629
2	3	Uttar Pradesh	1,584,764	1,687,818	8.39%	240.726
3	4	Gujarat	1,502,899	-	7.96%	228.290
4	5	Karnataka	1,493,127	1,631,977	7.91%	226.806
5	6	West Bengal	1,089,898	1,253,832	5.77%	165.556
6	7	Rajasthan	942,586	1,020,989	4.99%	143.179
7	8	Andhra Pradesh	862,957	972,782	4.57%	131.083
8	9	Telangana	861,031	969,604	4.56%	130.791
9	10	Madhya Pradesh	809,592	906,672	4.29%	122.977
10	11	Kerala	781,653	-	4.14%	118.733
11	12	Delhi	774,870	856,112	4.10%	117.703
12	13	Haryana	734,163	831,610	3.89%	111.519
13	14	Bihar	530,363	611,804	2.81%	80.562
14	15	Punjab	526,376	574,760	2.79%	79.957
15	16	Odisha	487,805	521,275	2.58%	74.098
16	17	Assam	315,881	-	1.67%	47.982
17	18	Chhattisgarh	304,063	329,180	1.61%	46.187
18	19	Jharkhand	297,204	328,598	1.57%	45.145
19	20	Uttarakhand	245,895	-	1.30%	37.351
20	21	Jammu & Kashmir	155,956	-	0.83%	23.690
21	22	Himachal Pradesh	153,845	165,472	0.81%	23.369
22	23	Goa	73,170	80,449	0.39%	11.115
23	24	Tripura	49,845	55,984	0.26%	7.571
24	25	Chandigarh	42,114	-	0.22%	6.397
25	26	Puducherry	34,433	38,253	0.18%	5.230

	Rank	State	GSDP(18-19)	GSDP(19-20)	Share(18-19)	GDP($Billion)
26	27	Meghalaya	33,481	36,572	0.18%	5.086
27	28	Sikkim	28,723	32,496	0.15%	4.363
28	29	Manipur	27,870	31,790	0.15%	4.233
29	30	Nagaland	27,283	-	0.14%	4.144
30	31	Arunachal Pradesh	24,603	-	0.13%	3.737
31	32	Mizoram	22,287	26,503	0.12%	3.385
32	33	Andaman & Nicobar Islands	-	-	-	-

In [19]:
```
driver.close()
```

Question 4: Scrape the details of trending repositories on Github.com. Url = https://github.com/ (https://github.com/)

In [20]:
```
driver = webdriver.Chrome()
driver.get("https://github.com/")
```

In [22]:
```
try:
    source = driver.find_element(By.XPATH,'//ul[@class="d-lg-flex list-style-none"]//li[3]//button[1]')
    source.click()
    trending = driver.find_element(By.XPATH,'//li[@class="HeaderMenu-item position-relative flex-wrap flex-ju
    trending.click()
except NoSuchElementException as e:

    print("Exception Raised: ", e)
```

In [25]:
```
title =[]
description=[]
count = []
language =[]

try:
    Title = driver.find_elements(By.XPATH,'//article[@class="Box-row"]//h2//a')
    for i in Title:
        title.append(i.text)
except NoSuchElementException:
    title.append('-')
try:
    Descri = driver.find_elements(By.XPATH,'//article[@class="Box-row"]//p')
    for i in Descri:
        description.append(i.text)
except NoSuchElementException:
    description.append('-')

try:
    Count  = driver.find_elements(By.XPATH,'//article[@class="Box-row"]//div[2]//a[2]')
    for i in Count:
        count.append(i.text)
except NoSuchElementException:
    count.append('-')

try:
    Lang_used = driver.find_elements(By.XPATH,'//article[@class="Box-row"]//div[2]//span[1]//span[2]')
    for i in Lang_used:
        language.append(i.text)
except NoSuchElementException:
    language.append('-')
```

In [26]:
```
print(len(title),len(description),len(count),len(language))
```

```
25 22 49 22
```

In [27]:
```
df = pd.DataFrame({'Title':title})
df1 = pd.DataFrame({'Description':description})
df2 = pd.DataFrame({'Contributors Count':count})
df3 = pd.DataFrame({'Language Used':language})
df4 = pd.concat([df,df1,df2,df3],axis=1)
df4.head(25)
```

Out[27]:

	Title	Description	Contributors Count	Language Used
0	LouisShark / chatgpt_system_prompt	store all agent's system prompt	735	C
1	federico-busato / Modern-CPP-Programming	Modern C++ Programming Course (C++11/14/17/20)		C++
2	nlohmann / json	JSON for Modern C++	339	Python
3	epfLLM / meditron	Meditron is a suite of open-source medical Lar...		Python
4	comfyanonymous / ComfyUI	The most powerful and modular stable diffusion...	6,454	Python
5	openai / openai-python	The official Python library for the OpenAI API		Go
6	grpc-ecosystem / grpc-gateway	gRPC to JSON proxy generator following the gRP...	24	Jupyter Notebook
7	linexjlin / GPTs	leaked prompts of GPTs		TypeScript
8	mlabonne / llm-course	Course with a roadmap and notebooks to get int...	1,628	TypeScript
9	kamranahmedse / developer-roadmap	Interactive roadmaps, guides and other educati...		TypeScript
10	makeplane / plane	🔥🔥🔥 Open Source JIRA, Linear and Height Alte...	2,143	Python
11	StanGirard / quivr	Your GenAI Second Brain 🧠 A personal productiv...		CSS
12	lllyasviel / Fooocus	Focus on prompting and generating	2,149	Python
13	straight-tamago / misaka	Effective prompting for Large Multimodal Model...		Swift
14	roboflow / multimodal-maestro	A library for building applications in a consi...	1,209	C
15	pointfreeco / swift-composable-architecture	Jailed iOS app that can install IPAs permanent...		Python
16	opa334 / TrollStore	分享 GitHub 上有趣、入门级的开源项目。Share interesting, entr...	192	MDX
17	521xueweihan / HelloGitHub	Examples and guides for using the OpenAI API	35,671	Ruby
18	openai / openai-cookbook	🚀 The easiest way to automate building and rel...		Jupyter Notebook
19	fastlane / fastlane	Ask Questions in natural language and get Answ...	890	Python
20	Vaibhavs10 / insanely-fast-whisper	Create Custom GPT and add/embed on your site u...		JavaScript
21	Ftindy / IPTV-URL	前端精读周刊。帮你理解最前沿、实用的技术。	2,730	JavaScript
22	danswer-ai / danswer	NaN		NaN

	Title	Description	Contributors Count	Language Used
23	SamurAIGPT / Open-Custom-GPT	NaN	1,597	NaN
24	ascoders / weekly	NaN		NaN

In [28]:
```
driver.close()
```

Question 5:Scrape the details of top 100 songs on billiboard.com. Url = https:/www.billboard.com/ (http://www.billboard.com/) You have to find the following details:

In [34]:
```
driver = webdriver.Chrome()
driver.get("http://www.billboard.com/")
```

In [35]:
```
try:
    charts = driver.find_element(By.XPATH,'//div[@class="lrv-u-flex lrv-u-flex-grow-1 lrv-u-flex-basis-60p"][
    charts.click()
except NoSuchElementException as e:

    print("Exception Raised: ", e)
```

In [37]:
```
try:
    top = driver.find_element(By.XPATH,'//a[@class="c-link  lrv-a-unstyle-link lrv-u-background-color-brand-s
    top.click()
except ElementClickInterceptedException as ec:
    print("Exception Raised: ", ec)
```

In [38]:
```
song =[]
artist=[]
last = []
peak =[]
wob = []
try:
    Title = driver.find_elements(By.XPATH,'//li[@class="lrv-u-width-100p"]//ul//li//h3')
    for i in Title:
        song.append(i.text)
except NoSuchElementException:
    song.append('-')
try:
    maker = driver.find_elements(By.XPATH,'//ul[@class="lrv-a-unstyle-list lrv-u-flex lrv-u-height-100p lrv-u
    for i in maker:
        artist.append(i.text)
except NoSuchElementException:
    artist.append('-')

try:
    LRank  = driver.find_elements(By.XPATH,'//div[@class="o-chart-results-list-row-container"]//ul//li[4]//ul
    for i in LRank:
        last.append(i.text)
except NoSuchElementException:
    last.append('-')

try:
    P_rank = driver.find_elements(By.XPATH,'//div[@class="o-chart-results-list-row-container"]//ul//li[4]//ul
    for i in P_rank:
        peak.append(i.text)
except NoSuchElementException:
    peak.append('-')

try:
    Weak = driver.find_elements(By.XPATH,'//div[@class="o-chart-results-list-row-container"]//ul//li[4]//ul//
    for i in Weak:
        wob.append(i.text)
except NoSuchElementException:
    wob.append('-')
```

In [39]:
```
print(len(song),len(artist),len(last),len(peak),len(wob))
```

```
100 100 200 200 100
```

In [40]:
```
df = pd.DataFrame({'Song Name':song,"Artist Name":artist,'Weak On Board':wob})
df1 = pd.DataFrame({'Last Weak Rank':last,'Peak Rank':peak})
df2 = pd.concat([df,df1],axis=1)
df2
```

Out[40]:

	Song Name	Artist Name	Weak On Board	Last Weak Rank	Peak Rank
0	Lovin On Me	Jack Harlow	2	2	1
1	Cruel Summer	Taylor Swift	29		
2	Paint The Town Red	Doja Cat	16	1	1
3	All I Want For Christmas Is You	Mariah Carey	60		
4	Snooze	SZA	50	3	1
...	...	...	...	...	...
195	NaN	NaN	NaN		
196	NaN	NaN	NaN	-	5
197	NaN	NaN	NaN		
198	NaN	NaN	NaN	96	22
199	NaN	NaN	NaN		

200 rows × 5 columns

In [41]:
```
driver.close()
```

Question 6: Scrape the details of Highest selling novels. Url - https://www.theguardian.com/news/datablog/2012/aug/09/best-

selling-books-all-time-fifty-shades-grey-compare (https://www.theguardian.com/news/datablog/2012/aug/09/best-selling-books-all-time-fifty-shades-grey-compare)

In [42]:
```
driver = webdriver.Chrome()
driver.get("https://www.theguardian.com/news/datablog/2012/aug/09/best-selling-books-all-time-fifty-shades-gr
```

In [43]:
```
book =[]
author=[]
sold = []
publisher =[]
genre = []
try:
    Title = driver.find_elements(By.XPATH,'//div[@class="embed block"]//table//tbody//tr//td[2]')
    for i in Title:
        book.append(i.text)
except NoSuchElementException:
    book.append('-')
try:
    writer = driver.find_elements(By.XPATH,'//div[@class="embed block"]//table//tbody//tr//td[3]')
    for i in writer:
        author.append(i.text)
except NoSuchElementException:
    author.append('-')

try:
    Volumes  = driver.find_elements(By.XPATH,'//div[@class="embed block"]//table//tbody//tr//td[4]')
    for i in Volumes:
        sold.append(i.text)
except NoSuchElementException:
    sold.append('-')

try:
    Publish = driver.find_elements(By.XPATH,'//div[@class="embed block"]//table//tbody//tr//td[5]')
    for i in Publish:
        publisher.append(i.text)
except NoSuchElementException:
    publisher.append('-')

try:
    Genre = driver.find_elements(By.XPATH,'//div[@class="embed block"]//table//tbody//tr//td[6]')
    for i in Genre:
        genre.append(i.text)
except NoSuchElementException:
    genre.append('-')
```

In [44]:
```
df = pd.DataFrame({"Book Name":book,"Author Name":author,"Volumes Sold":sold,"Publisher":publisher,"Genre":ge
df
```

Out[44]:

	Book Name	Author Name	Volumes Sold	Publisher	Genre
0	Da Vinci Code,The	Brown, Dan	5,094,805	Transworld	Crime, Thriller & Adventure
1	Harry Potter and the Deathly Hallows	Rowling, J.K.	4,475,152	Bloomsbury	Children's Fiction
2	Harry Potter and the Philosopher's Stone	Rowling, J.K.	4,200,654	Bloomsbury	Children's Fiction
3	Harry Potter and the Order of the Phoenix	Rowling, J.K.	4,179,479	Bloomsbury	Children's Fiction
4	Fifty Shades of Grey	James, E. L.	3,758,936	Random House	Romance & Sagas
...	...	...	...	...	...
95	Ghost,The	Harris, Robert	807,311	Random House	General & Literary Fiction
96	Happy Days with the Naked Chef	Oliver, Jamie	794,201	Penguin	Food & Drink: General
97	Hunger Games,The:Hunger Games Trilogy	Collins, Suzanne	792,187	Scholastic Ltd.	Young Adult Fiction
98	Lost Boy,The:A Foster Child's Search for the L...	Pelzer, Dave	791,507	Orion	Biography: General
99	Jamie's Ministry of Food:Anyone Can Learn to C...	Oliver, Jamie	791,095	Penguin	Food & Drink: General

100 rows × 5 columns

In [45]:
```
driver.close()
```

Question 7:Scrape the details most watched tv series of all time from imdb.com. Url = https://www.imdb.com/list/ls095964455/ (https://www.imdb.com/list/ls095964455/) You have to find the following details:

In [46]:
```
driver = webdriver.Chrome()
driver.get("https://www.imdb.com/list/ls095964455/")
```

In [47]:
```
movie =[]
year=[]
time = []
rating =[]
genre = []
votes = []
try:
    Title = driver.find_elements(By.XPATH,'//div[@class="lister-item-content"]//h3//a')
    for i in Title:
        movie.append(i.text)
except NoSuchElementException:
    movie.append('-')
try:
    Y_span = driver.find_elements(By.XPATH,'//div[@class="lister-item-content"]//h3//span[2]')
    for i in Y_span:
        year.append(i.text)
except NoSuchElementException:
    year.append('-')

try:
    R_time  = driver.find_elements(By.XPATH,'//div[@class="lister-item-content"]//p//span[3]')
    for i in R_time:
        time.append(i.text)
except NoSuchElementException:
    time.append('-')

try:
    Rate = driver.find_elements(By.XPATH,'//div[@class="ipl-rating-star small"]//span[2]')
    for i in Rate:
        rating.append(i.text)
except NoSuchElementException:
    rating.append('-')

try:
    Genre = driver.find_elements(By.XPATH,'//div[@class="lister-item-content"]//p//span[5]')
    for i in Genre:
        genre.append(i.text)
except NoSuchElementException:
    genre.append('-')

try:
    Vote = driver.find_elements(By.XPATH,'//div[@class="lister-item-content"]//p[4]//span[2]')
    for i in Vote:
```

```
        votes.append(i.text)
except NoSuchElementException:
    votes.append('-')
```

In [48]:
```
df = pd.DataFrame({"Movie Name":movie,"Year Span":year,"Run Time":time,"Ratings":rating,"Genre":genre,"Votes"
df
```

Out[48]:

	Movie Name	Year Span	Run Time	Ratings	Genre	Votes
0	Game of Thrones	(2011–2019)	4,189 min	9.2	Action, Adventure, Drama	2,226,676
1	Stranger Things	(2016–2025)	51 min	8.7	Drama, Fantasy, Horror	1,293,610
2	The Walking Dead	(2010–2022)	44 min	8.1	Drama, Horror, Thriller	1,056,272
3	13 Reasons Why	(2017–2020)	60 min	7.5	Drama, Mystery, Thriller	310,129
4	The 100	(2014–2020)	43 min	7.6	Drama, Mystery, Sci-Fi	269,253
...	...	...	...	...	...	...
95	Reign	(2013–2017)	42 min	7.5	Drama	53,278
96	A Series of Unfortunate Events	(2017–2019)	50 min	7.8	Adventure, Comedy, Drama	65,358
97	Criminal Minds	(2005–)	42 min	8.1	Crime, Drama, Mystery	213,175
98	Scream: The TV Series	(2015–2019)	45 min	7	Comedy, Crime, Drama	44,372
99	The Haunting of Hill House	(2018)	572 min	8.6	Drama, Horror, Mystery	278,065

100 rows × 6 columns

In [49]:
```
driver.close()
```

Question 8: Details of Datasets from UCI machine learning repositories. Url = https://archive.ics.uci.edu/ (https://archive.ics.uci.edu/) You have to find the following details:

In [50]:
```
driver = webdriver.Chrome()
driver.get("https://archive.ics.uci.edu/")
```

In [51]:
```
All = driver.find_element(By.XPATH,'//a[@class="btn-primary btn"]')
All.click()
```

In [52]:
```
try:
    expand = driver.find_element(By.XPATH,'//select[@class="select-primary select select-sm rounded-full"]//o
    expand.click()
except ElementClickInterceptedException as e:
    print("Exception Raised : ",e)
```

In [53]:
```
expand1 = driver.find_element(By.XPATH,'//div[@class="flex flex-wrap items-center gap-4"]//label[2]//div[2]')
expand1.click()
```

In [54]:
```
name = []
Type=[]
task=[]
f_type=[]
instances=[]
attribute =[]
year =[]
Name = driver.find_elements(By.XPATH,'//div[@class="rounded-box bg-base-100"]//div//div[2]//h2//a')
for i in Name:
    name.append(i.text)
data = driver.find_elements(By.XPATH,'//div[@class="rounded-box bg-base-100"]//div//div[2]//div//div[2]//span
for i in data:
    Type.append(i.text)
Task = driver.find_elements(By.XPATH,'//div[@class="rounded-box bg-base-100"]//div//div[2]//div//div[1]//span
for i in Task:
    task.append(i.text)
feature= driver.find_elements(By.XPATH,'//div[@class="grid grid-cols-8 overflow-x-auto"]//table//tbody//tr//t
for i in feature:
    f_type.append(i)
Instance = driver.find_elements(By.XPATH,'//div[@class="rounded-box bg-base-100"]//div//div[2]//div//div[3]//
for i in Instance:
    instances.append(i.text)
attri = driver.find_elements(By.XPATH,'//div[@class="rounded-box bg-base-100"]//div//div[2]//div//div[4]//spa
for i in attri:
    attribute.append(i.text)
Year = driver.find_elements(By.XPATH,'//div[@class="grid grid-cols-8 overflow-x-auto"]//table//tbody//tr//td[
for i in Year:
    year.append(i.text)
```

In [55]:
```
print(len(name),len(Type),len(task),len(f_type),len(instances),len(attribute),len(year))
```

```
25 25 25 25 25 25 25
```

In [56]:
```
df= pd.DataFrame({'Dataset Name':name,'Data Type':Type,'Task':task,"Attribute Type":f_type,'No. Of Instances'
df
```

Out[56]:

	Dataset Name	Data Type	Task	Attribute Type	No. Of Instances	No. Of Attributes	Year
0	Iris	Tabular	Classification	<selenium.webdriver.remote.webelement.WebEleme...	150 Instances	4 Features	7/1/1988
1	Heart Disease	Multivariate	Classification	<selenium.webdriver.remote.webelement.WebEleme...	303 Instances	13 Features	7/1/1988
2	Adult	Multivariate	Classification	<selenium.webdriver.remote.webelement.WebEleme...	48.84K Instances	14 Features	5/1/1996
3	Wine	Tabular	Classification	<selenium.webdriver.remote.webelement.WebEleme...	178 Instances	13 Features	7/1/1991
4	Breast Cancer Wisconsin (Diagnostic)	Multivariate	Classification	<selenium.webdriver.remote.webelement.WebEleme...	569 Instances	30 Features	11/1/1995
5	Diabetes	Multivariate, Time-Series	Classification	<selenium.webdriver.remote.webelement.WebEleme...	1 Instances	20 Features	N/A
6	Dry Bean Dataset	Multivariate	Classification	<selenium.webdriver.remote.webelement.WebEleme...	13.61K Instances	16 Features	9/14/2020
7	Car Evaluation	Multivariate	Classification	<selenium.webdriver.remote.webelement.WebEleme...	1.73K Instances	6 Features	6/1/1997
8	Wine Quality	Multivariate	Classification, Regression	<selenium.webdriver.remote.webelement.WebEleme...	4.9K Instances	12 Features	10/7/2009
9	Bank Marketing	Multivariate	Classification	<selenium.webdriver.remote.webelement.WebEleme...	45.21K Instances	17 Features	2/14/2012
10	Mushroom	Multivariate	Classification	<selenium.webdriver.remote.webelement.WebEleme...	8.12K Instances	22 Features	4/27/1987
11	Rice (Cammeo and Osmancik)	Multivariate	Classification	<selenium.webdriver.remote.webelement.WebEleme...	3.81K Instances	7 Features	10/6/2019
12	Abalone	Tabular	Classification, Regression	<selenium.webdriver.remote.webelement.WebEleme...	4.18K Instances	8 Features	12/1/1995
13	Census Income	Multivariate	Classification	<selenium.webdriver.remote.webelement.WebEleme...	48.84K Instances	14 Features	5/1/1996
14	Student Performance	Multivariate	Classification, Regression	<selenium.webdriver.remote.webelement.WebEleme...	649 Instances	33 Features	11/27/2014
15	Statlog (German Credit Data)	Multivariate	Classification	<selenium.webdriver.remote.webelement.WebEleme...	1K Instances	20 Features	11/17/1994

	Dataset Name	Data Type	Task	Attribute Type	No. Of Instances	No. Of Attributes	Year
16	Automobile	Multivariate	Regression	<selenium.webdriver.remote.webelement.WebEleme...	205 Instances	25 Features	5/19/1987
17	Breast Cancer	Multivariate	Classification	<selenium.webdriver.remote.webelement.WebEleme...	286 Instances	9 Features	7/11/1988
18	Auto MPG	Multivariate	Regression	<selenium.webdriver.remote.webelement.WebEleme...	398 Instances	7 Features	7/7/1993
19	Breast Cancer Wisconsin (Original)	Multivariate	Classification	<selenium.webdriver.remote.webelement.WebEleme...	699 Instances	9 Features	7/15/1992
20	Online Retail	Multivariate, Sequential, Time-Series	Classification, Clustering	<selenium.webdriver.remote.webelement.WebEleme...	541.91K Instances	8 Features	11/6/2015
21	Predict students' dropout and academic success	Tabular	Classification	<selenium.webdriver.remote.webelement.WebEleme...	4.42K Instances	36 Features	12/13/2021
22	Spambase	Multivariate	Classification	<selenium.webdriver.remote.webelement.WebEleme...	4.6K Instances	57 Features	7/1/1999
23	Glass Identification	Multivariate	Classification	<selenium.webdriver.remote.webelement.WebEleme...	214 Instances	9 Features	9/1/1987
24	Thyroid Disease	Multivariate, Domain-Theory	Classification	<selenium.webdriver.remote.webelement.WebEleme...	7.2K Instances	5 Features	1/1/1987

In [57]:
```
driver.close()
```

In [ ]:

In [ ]: