Web Scraping Assignment 4

In [1]:
```
# import debugger
import re

# selenium
import selenium
import pandas as pd
from selenium import webdriver

# Beautiful soup
from bs4 import BeautifulSoup
import requests

# add time
import time

from selenium.common.exceptions import NoSuchElementException, StaleElementReferenceException

from selenium.webdriver.support.ui import WebDriverWait
```

Q1 : Scrape the details of most viewed videos on YouTube from Wikipedia:

Url = https://en.wikipedia.org/wiki/List_of_most-viewed_YouTube_videos/
You need to find following details:
A) Rank
B) Name
C) Artist
D) Upload date
E) Views

In [2]:
```
# first, connect to the webdriver
driver=webdriver.Chrome(r"C:/Users/HP/Downloads/chromedriver_win32 (1)/chromedriver.exe")

# getting the webpage of mentioned url
url = 'https://en.wikipedia.org/wiki/List_of_most-viewed_YouTube_videos'
driver.get(url)
```

In [3]:
```
# creating empty list for scraping the data

Rank = []
Name = []
Artist = []
Date = []
Views = []
```

In [4]:
```
# scraping Rank of the videos
try:
    for i in driver.find_elements_by_xpath("//table[@class='wikitable sortable jquery-tablesorter'][1]/tbody/tr/td[1]"):
        Rank.append(i.text)
except NoSuchElementException:
    Rank.append("-")

# Scraping Name of the videos
try:
    for i in driver.find_elements_by_xpath("//table[@class='wikitable sortable jquery-tablesorter'][1]/tbody/tr/td[2]"):
        Name.append(i.text)
except NoSuchElementException:
    Name.append("-")

# Scraping Artist of the videos
try:
    for i in driver.find_elements_by_xpath("//table[@class='wikitable sortable jquery-tablesorter'][1]/tbody/tr/td[3]"):
        Artist.append(i.text)
except NoSuchElementException:
    Artist.append("-")

# Scraping Upload_Date of the videos
try:
    for i in driver.find_elements_by_xpath("//table[@class='wikitable sortable jquery-tablesorter'][1]/tbody/tr/td[5]"):
        Date.append(i.text)
except NoSuchElementException:
    Date.append("-")

# Scraping Views of the videos
try:
    for i in driver.find_elements_by_xpath("//table[@class='wikitable sortable jquery-tablesorter'][1]/tbody/tr/td[4]"):
        Views.append(i.text)
except NoSuchElementException:
    Views.append("-")

# creating DataFrame for scraped data
Wiki = pd.DataFrame({})
Wiki['Rank'] = Rank
Wiki['Name'] = Name
Wiki['Artist'] = Artist
Wiki['Upload Date'] = Date
Wiki['Views (in Billions)'] = Views

# removing stray numbers from Name column
Wiki.Name = Wiki.Name.apply(lambda x:x[:-4].strip('"'))
Wiki
```

Out[4]:

	Rank	Name	Artist	Upload Date	Views (in Billions)
0	1.	Baby Shark Dance	Pinkfong Baby Shark - Kids' Songs & Stories	June 17, 2016	9.58
1	2.	Despacito	Luis Fonsi	January 12, 2017	7.61

2	3.	Johny Johny Yes Papa	LooLoo Kids	October 8, 2016	5.85
3	4.	Shape of You	Ed Sheeran	January 30, 2017	5.51
4	5.	See You Again	Wiz Khalifa	April 6, 2015	5.31
5	6.	Bath Song	Cocomelon – Nursery Rhymes	May 2, 2018	4.60
6	7.	Learning Colors – Colorful Eggs on a Farm	Miroshka TV	February 27, 2018	4.53
7	8.	Masha and the Bear – Recipe for Disaster	Get Movies	January 31, 2012	4.47
8	9.	Uptown Funk	Mark Ronson	November 19, 2014	4.34
9	10.	Phonics Song with Two Words	ChuChu TV	March 6, 2014	4.27
10	11.	Gangnam Style	Psy	July 15, 2012	4.24
11	12.	Dame Tu Cosita	El Chombo	April 5, 2018	3.67
12	13.	Sugar	Maroon 5	January 14, 2015	3.58
13	14.	Sorry	Justin Bieber	October 22, 2015	3.48
14	15.	Roar	Katy Perry	September 5, 2013	3.46
15	16.	Counting Stars	OneRepublic	May 31, 2013	3.43
16	17.	Thinking Out Loud	Ed Sheeran	October 7, 2014	3.35
17	18.	Wheels on the Bus	Cocomelon – Nursery Rhymes	May 24, 2018	3.33
18	19.	Faded	Alan Walker	December 3, 2015	3.17
19	20.	Dark Horse	Katy Perry	February 20, 2014	3.17
20	21.	Girls Like You	Maroon 5	May 31, 2018	3.16
21	22.	Lean On	Major Lazer	March 22, 2015	3.12
22	23.	Bailando	Enrique Iglesias	April 11, 2014	3.12
23	24.	Shake It Off	Taylor Swift	August 18, 2014	3.11
24	25.	Let Her Go	Passenger	July 25, 2012	3.11
25	26.	Axel F	Crazy Frog	June 16, 2009	3.08
26	27.	Perfect	Ed Sheeran	November 9, 2017	3.00
27	28.	Mi Gente	J Balvin	June 29, 2017	3.00
28	29.	Waka Waka (This Time for Africa)	Shakira	June 4, 2010	2.98
29	30.	Hello	Adele	October 22, 2015	2.90

In [5]:
```
print(len(Rank),
len(Name),
len(Artist),
len(Date),
len(Views))
```

```
30 30 30 30 30
```

In [6]:
```
driver.close()
```

In [ ]:
```

```

Q2 : Scrape the details team India's international fixtures from bcci.tv.

```
Url = https://www.bcci.tv/.
You need to find following details:
 A) Match title (I.e. 1st ODI)
 B) Series
 C) Place
 D) Date
 E) Time

   Note: - From bcci.tv home page you have reach to the international fixture page through code.
```

In [2]:
```
# connecting to the webdriver
driver=webdriver.Chrome(r"C:/Users/HP/Downloads/chromedriver_win32 (1)/chromedriver.exe")

# getting the webpage of mentioned url
url=('https://www.bcci.tv/')
driver.get(url)
```

In [3]:
```
btn=driver.find_element_by_xpath("//div[@class='navigation__drop-down drop-down drop-down--reveal-on-hover']/div/ul/li/a")
driver.get(btn.get_attribute("href"))
time.sleep(3)

# creating empty lists for scraping the data
Match_Title = []
Series = []
Place = []
Date = []
Time = []
```

In [4]:
```
for i in driver.find_elements_by_xpath("//div[@class='fixture__format-strip']/span[@class='u-unskewed-text fixture__format']"):
    Match_Title.append(i.text)

for i in driver.find_elements_by_xpath("//div[@class='fixture__format-strip']/span[@class='u-unskewed-text fixture__tournament-label u-tru
    Series.append(i.text)

for i in driver.find_elements_by_xpath("//div[@class='fixture__description u-unskewed-text']/p/span"):
    Place.append(i.text)

for i in driver.find_elements_by_xpath("//span[@class='fixture__datetime tablet-only']/strong[1]"):
    Date.append(i.text.replace('\n',' '))

date=[i.split(' ',3)[:3] for i in Date]
```

```
date=[' '.join(i) for i in date]
Time=[i.split(' ',3)[-1] for i in Date]

# creating data frame
fixture=pd.DataFrame({'Match Title': Match_Title,
                      "Series": Series,
                      "Place": Place,
                      "Date": date,
                      "Time": Time})
fixture
```

Out[4]:

	Match Title	Series	Place	Date	Time
0	T20I	2021 ICC MEN'S T20 WORLD CUP	Dubai International Stadium, Dubai	Monday 8 November	November
1	T20I	INDIA V NEW ZEALAND 2021	Sawai Mansingh Stadium, Jaipur	Wednesday 17 November	November
2	T20I	INDIA V NEW ZEALAND 2021	JSCA International Stadium Complex, Ranchi	Friday 19 November	November
3	T20I	INDIA V NEW ZEALAND 2021	Eden Gardens, Kolkata	Sunday 21 November	November
4	TEST	INDIA V NEW ZEALAND 2021	Green Park, Kanpur	Thursday 25 November	November
5	TEST	INDIA V NEW ZEALAND 2021	Wankhede Stadium, Mumbai	Friday 3 December	December
6	ODI	INDIA V WEST INDIES 2022	Narendra Modi Stadium, Ahmedabad	Sunday 6 February	February
7	ODI	INDIA V WEST INDIES 2022	Sawai Mansingh Stadium, Jaipur	Wednesday 9 February	February
8	ODI	INDIA V WEST INDIES 2022	Eden Gardens, Kolkata	Saturday 12 February	February
9	T20I	INDIA V WEST INDIES 2022	Barabati Stadium, Cuttack	Tuesday 15 February	February
10	T20I	INDIA V WEST INDIES 2022	ACA-VDCA Stadium, Visakhapatnam	Friday 18 February	February
11	T20I	INDIA V WEST INDIES 2022	Greenfield Stadium, Thiruvananthapuram	Sunday 20 February	February
12	TEST	INDIA V SRI LANKA 2022	M.Chinnaswamy Stadium, Bangalore	Friday 25 February	February
13	TEST	INDIA V SRI LANKA 2022	Punjab Cricket Association Stadium, Mohali, Ch...	Saturday 5 March	March
14	T20I	INDIA V SRI LANKA 2022	Punjab Cricket Association Stadium, Mohali, Ch...	Sunday 13 March	March
15	T20I	INDIA V SRI LANKA 2022	Himachal Pradesh Cricket Association Stadium, ...	Tuesday 15 March	March
16	T20I	INDIA V SRI LANKA 2022	Bharat Ratna Shri Atal Bihari Vajpai Ekana Cri...	Friday 18 March	March
17	T20I	INDIA V SOUTH AFRICA 2022	M. A. Chidambaram Stadium, Chennai	Thursday 9 June	June
18	T20I	INDIA V SOUTH AFRICA 2022	M.Chinnaswamy Stadium, Bangalore	Sunday 12 June	June
19	T20I	INDIA V SOUTH AFRICA 2022	Vidarbha Cricket Association Stadium, Nagpur	Tuesday 14 June	June

In [5]:
```
len(url)
```

Out[5]:
```
20
```

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

In [ ]:

In [ ]:

Q 3: Scrape the details of selenium exception from guru99.com.

```
Url = https://www.guru99.com/
You need to find following details:
A) Name
B) Description
Note: - From guru99 home page you have to reach to selenium exception handling page through code.
```

In [2]:
```
# connecting to the web driver
driver=webdriver.Chrome(r"C:/Users/HP/Downloads/chromedriver_win32 (1)/chromedriver.exe")

# getting the webpage of mentioned url
url = ("https://www.guru99.com/")
driver.get(url)
```

In [3]:
```
# creating empty list
Name = []
Description = []
```

In [4]:
```
# clicking on Selenium button
driver.find_element_by_xpath("//li//a[@title='Selenium']").click()

# clicking on Exception Handling button
driver.find_element_by_xpath('//a[@title="Selenium Exception Handling (Common Exceptions List)"]').click()

# scraping Name
for i in driver.find_elements_by_xpath("//table[@class='table table-striped']/tbody/tr/td[1]"):
    Name.append(i.text)

# scraping Description
for i in driver.find_elements_by_xpath("//table[@class='table table-striped']/tbody/tr/td[2]"):
    Description.append(i.text)


# creating the dataframe from the scraped data
Selenium = pd.DataFrame({})
Selenium['Exception_Name'] = Name
Selenium['Description'] = Description
Selenium
```

Out[4]:

	Exception_Name	Description
0	ElementNotVisibleException	This type of Selenium exception occurs when an...

1	ElementNotSelectableException	This Selenium exception occurs when an element...
2	NoSuchElementException	This Exception occurs if an element could not ...
3	NoSuchFrameException	This Exception occurs if the frame target to b...
4	NoAlertPresentException	This Exception occurs when you switch to no pr...
5	NoSuchWindowException	This Exception occurs if the window target to ...
6	StaleElementReferenceException	This Selenium exception occurs happens when th...
7	SessionNotFoundException	The WebDriver is acting after you quit the bro...
8	TimeoutException	Thrown when there is not enough time for a com...
9	WebDriverException	This Exception takes place when the WebDriver ...
10	ConnectionClosedException	This type of Exception takes place when there ...
11	ElementClickInterceptedException	The command may not be completed as the elemen...
12	ElementNotInteractableException	This Selenium exception is thrown when any ele...
13	ErrorInResponseException	This happens while interacting with the Firefo...
14	ErrorHandler.UnknownServerException	Exception is used as a placeholder in case if ...
15	ImeActivationFailedException	This expectation will occur when IME engine ac...
16	ImeNotAvailableException	It takes place when IME support is unavailable.
17	InsecureCertificateException	Navigation made the user agent to hit a certif...
18	InvalidArgumentException	It occurs when an argument does not belong to ...
19	InvalidCookieDomainException	This happens when you try to add a cookie unde...
20	InvalidCoordinatesException	This type of Exception matches an interacting ...
21	InvalidElementStateExceptio	It occurs when command can't be finished when ...
22	InvalidSessionIdException	This Exception took place when the given sessi...
23	InvalidSwitchToTargetException	This occurs when the frame or window target to...
24	JavascriptException	This issue occurs while executing JavaScript g...
25	JsonException	It occurs when you afford to get the session w...
26	NoSuchAttributeException	This kind of Exception occurs when the attribu...
27	MoveTargetOutOfBoundsException	It takes place if the target provided to the A...
28	NoSuchContextException	ContextAware does mobile device testing.
29	NoSuchCookieException	This Exception occurs when no cookie matching ...
30	NotFoundException	This Exception is a subclass of WebDriverExcep...
31	RemoteDriverServerException	This Selenium exception is thrown when the ser...
32	ScreenshotException	It is not possible to capture a screen.
33	SessionNotCreatedException	It happens when a new session could not be suc...
34	UnableToSetCookieException	This occurs if a driver is unable to set a coo...
35	UnexpectedTagNameException	Happens if a support class did not get a web e...
36	UnhandledAlertException	This expectation occurs when there is an alert...
37	UnexpectedAlertPresentException	It occurs when there is the appearance of an u...
38	UnknownMethodException	This Exception happens when the requested comm...
39	UnreachableBrowserException	This Exception occurs only when the browser is...
40	UnsupportedCommandException	This occurs when remote WebDriver does n't sen...

In [5]:
```
driver.close()
```

In [ ]:

In [ ]:

Q 4 : Scrape the details of State-wise GDP of India from statisticstime.com.

```
Url = http://statisticstimes.com/
You have to find following details:
A) Rank
B) State
C) GSDP at current price (19-20)
D) GSDP at current price (18-19)
E) Share(18-19)
F) GDP($ billion)
  Note: - From statisticstimes home page you have to reach to economy page through code.
```

In [6]:
```
# connecting to the web driver
driver=webdriver.Chrome(r"C:/Users/HP/Downloads/chromedriver_win32 (1)/chromedriver.exe")

# getting the webpage of mentioned url
url = ("https://statisticstimes.com/")
driver.get(url)
```

In [8]:
```
# clicking on Economy button
driver.find_element_by_xpath("//div[@class='navbar']/div[2]/button").click()

# clicking on India
driver.find_element_by_xpath("//div[@class='dropdown-content']/a[3]").click()
time.sleep(3)

# clicking on GDP of Indian Economy
GDP = driver.find_element_by_xpath("/html/body/div[2]/div[2]/div[2]/ul/li[1]/a").click()
time.sleep(3)
```

In [9]:

```
# creating empty list
Rank = []
State = []
GSDP1 = []
GSDP2 = []
Share = []
GDP_billion = []

# scraping Rank
try:
    for i in driver.find_elements_by_xpath("//table[@class='display dataTable']/tbody/tr/td[1]"):
        Rank.append(i.text)
except NoSuchElementException:
    Rank.append("_")

# scraping State
try:
    for i in driver.find_elements_by_xpath("//table[@class='display dataTable']/tbody/tr/td[2]"):
        State.append(i.text)
except NoSuchElementException:
    State.append("_")

# scraping GSDP at current price (19-20)
try:
    for i in driver.find_elements_by_xpath("//table[@class='display dataTable']/tbody/tr/td[3]"):
        GSDP1.append(i.text)
except NoSuchElementException:
    GSDP1.append("_")

# scraping GSDP at current price (18-19)
try:
    for i in driver.find_elements_by_xpath("//table[@class='display dataTable']/tbody/tr/td[4]"):
        GSDP2.append(i.text)
except NoSuchElementException:
    GSDP2.append("_")

# scraping Share (18-19)
try:
    for i in driver.find_elements_by_xpath("//table[@class='display dataTable']/tbody/tr/td[5]"):
        Share.append(i.text)
except NoSuchElementException:
    Share.append("_")

# scraping GDP $ billion
try:
    for i in driver.find_elements_by_xpath("//table[@class='display dataTable']/tbody/tr/td[6]"):
        GDP_billion.append(i.text)
except NoSuchElementException:
    GDP_billion.append("_")


# creating DataFrame from the scraped data
GDP = pd.DataFrame({})
GDP['Rank'] = Rank
GDP['State'] = State
GDP['GSDP at current price (19-20)'] = GSDP1
GDP['GSDP at current price (18-19)'] = GSDP2
GDP['Share (18-19)'] = Share
GDP['GDP($ billion)'] = GDP_billion
GDP
```

Out[9]:

	Rank	State	GSDP at current price (19-20)	GSDP at current price (18-19)	Share (18-19)	GDP($ billion)
0	1	Maharashtra	-	2,632,792	13.94%	399.921
1	2	Tamil Nadu	1,845,853	1,630,208	8.63%	247.629
2	3	Uttar Pradesh	1,687,818	1,584,764	8.39%	240.726
3	4	Gujarat	-	1,502,899	7.96%	228.290
4	5	Karnataka	1,631,977	1,493,127	7.91%	226.806
...	...	...	...	...	...	...
61	29	Sikkim	28,391	25,141	0.15%	17,060
62	30	Nagaland	-	24,534	0.15%	-
63	31	Arunachal Pradesh	-	22,488	0.13%	-
64	32	Mizoram	24,424	20,947	0.13%	17,797
65	33	Andaman & Nicobar Islands	-	-	-	-

66 rows × 6 columns

In [10]:

```
driver.close()
```

In [ ]:

In [ ]:

Q 5 : Scrape the details of trending repositories on Github.com.

```
Url = https://github.com/
You have to find the following details:
A) Repository title
B) Repository description
C) Contributors count
D) Language used
  Note: - From the home page you have to click on the trending option from Explore menu through code.
```

In [2]:
```
# connecting to the web driver
driver=webdriver.Chrome(r"C:/Users/HP/Downloads/chromedriver_win32 (1)/chromedriver.exe")

# getting the webpage of mentioned url
url = ("https://github.com/")
driver.get(url)
```

In [3]:
```
# getting explore button and clicking on it
explore = driver.find_element_by_xpath("/html/body/div[1]/header/div/div[2]/nav/ul/li[4]/details").click()

# selecting trending option
trend_url = driver.find_element_by_xpath("/html/body/div[1]/header/div/div[2]/nav/ul/li[4]/details/div/ul[2]/li[3]/a")
urls = trend_url.get_attribute("href")
driver.get(urls)
```

In [4]:
```
# creating empty list
URLs = []
repository_title = []
Description = []
Contributors = []
Language = []
lang = []

# fetching urls for each repository
repository = driver.find_elements_by_xpath("//h1[@class='h3 lh-condensed']//a")
for i in repository:
    URLs.append(i.get_attribute("href"))

# scraping Repository title data
title = driver.find_elements_by_xpath("//h1[@class = 'h3 lh-condensed']")
for i in title:
    repository_title.append(i.text)

# scraping data from all repository page
for i in URLs:
    driver.get(i)
    time.sleep(5)

    # scraping Repository Description data
    try:
        desc = driver.find_element_by_xpath("//p[@class='f4 mt-3']")
        Description.append(desc.text)
    except NoSuchElementException:
        Description.append('-')


    # scraping Contributors Count data
    try:
        contributor = driver.find_element_by_xpath("//*[contains(text(),'    Contributors ')]")
        Contributors.append(contributor.text.replace('Contributors',''))
    except NoSuchElementException:
        Contributors.append('-')


    # scraping Languages used data
    try:
        for i in driver.find_elements_by_xpath("//ul[@class= 'list-style-none']//li//span[1]"):
            lang.append(i.text)
        Language.append(lang)
    except NoSuchElementException:
        Language.append('-')


# Data Framing
Github = pd.DataFrame({})
Github['Repository Title'] = repository_title
Github['Repository Description'] = Description
Github['Contributors Count'] = Contributors
Github['Language Used'] = Language
Github
```

Out[4]:

	Repository Title	Repository Description	Contributors Count	Language Used
0	misterokaygo / MapAssist	-	16	[C#, Java, C++, C, Assembly, TypeScript, JavaS...
1	questdb / questdb	An open source SQL database designed to proces...	50	[C#, Java, C++, C, Assembly, TypeScript, JavaS...
2	babysor / MockingBird	🚀AI拟声: 5秒内克隆您的声音并生成任意语音内容 Clone a voice in 5 s...	13	[C#, Java, C++, C, Assembly, TypeScript, JavaS...
3	v2fly / v2ray-core	A platform for building proxies to bypass netw...	111	[C#, Java, C++, C, Assembly, TypeScript, JavaS...
4	foxsen / archbase	教科书《计算机体系结构基础》（胡伟武等，第三版）的开源版本	-	[C#, Java, C++, C, Assembly, TypeScript, JavaS...
5	DrKLO / Telegram	Telegram for Android source	16	[C#, Java, C++, C, Assembly, TypeScript, JavaS...
6	files-community / Files	A modern file manager that pushes the boundari...	144	[C#, Java, C++, C, Assembly, TypeScript, JavaS...
7	Dreamacro / clash	A rule-based tunnel in Go.	67	[C#, Java, C++, C, Assembly, TypeScript, JavaS...
8	vandadnp / flutter-tips-and-tricks	A Collection of Flutter and Dart Tips and Tricks	-	[C#, Java, C++, C, Assembly, TypeScript, JavaS...
9	233boy / v2ray	最好用的 V2Ray 一键安装脚本 & 管理脚本	-	[C#, Java, C++, C, Assembly, TypeScript, JavaS...
10	aeon0 / botty	D2R Pixel Bot	2	[C#, Java, C++, C, Assembly, TypeScript, JavaS...
11	commaai / openpilot	openpilot is an open source driver assistance ...	254	[C#, Java, C++, C, Assembly, TypeScript, JavaS...
12	jwasham / coding-interview-university	A complete computer science study plan to beco...	226	[C#, Java, C++, C, Assembly, TypeScript, JavaS...
13	ossu / computer-science	🎓 Path to a free self-taught education in Comp...	116	[C#, Java, C++, C, Assembly, TypeScript, JavaS...

14	bitcoin / bitcoin	Bitcoin Core integration/staging tree	828	[C#, Java, C++, C, Assembly, TypeScript, JavaS...
15	solana-labs / solana	Web-Scale Blockchain for fast, secure, scalabl...	223	[C#, Java, C++, C, Assembly, TypeScript, JavaS...
16	getfotiaoqiang / download	佛跳墙官方版本下载页 翻墙 代理 科学上网 外网 加速器 梯子 路由	-	[C#, Java, C++, C, Assembly, TypeScript, JavaS...
17	Kholid060 / automa	A chrome extension for automating your browser...	4	[C#, Java, C++, C, Assembly, TypeScript, JavaS...
18	dockersamples / example-voting-app	Example Docker Compose app	32	[C#, Java, C++, C, Assembly, TypeScript, JavaS...
19	sherlock-project / sherlock	🔎 Hunt down social media accounts by username ...	135	[C#, Java, C++, C, Assembly, TypeScript, JavaS...
20	ACL4SSR / ACL4SSR	SSR 去广告ACL规则/SS完整GFWList规则/Clash规则碎片，Telegram频...	-	[C#, Java, C++, C, Assembly, TypeScript, JavaS...
21	freefq / free	翻墙、免费翻墙、免费科学上网、免费节点、免费梯子、免费ss/v2ray/trojan节点、蓝...	2	[C#, Java, C++, C, Assembly, TypeScript, JavaS...
22	TandoorRecipes / recipes	Application for managing recipes, planning mea...	64	[C#, Java, C++, C, Assembly, TypeScript, JavaS...
23	CSSEGISandData / COVID-19	Novel Coronavirus (COVID-19) Cases, provided b...	8	[C#, Java, C++, C, Assembly, TypeScript, JavaS...
24	obsproject / obs-studio	OBS Studio - Free and open source software for...	461	[C#, Java, C++, C, Assembly, TypeScript, JavaS...

In [5]:
```
driver.close()
```

In [ ]:

In [ ]:

Q 6 : 6. Scrape the details of top 100 songs on billboard.com.

Url = https://www.billboard.com/
You have to find the following details:
A) Song name
B) Artist name
C) Last week rank
D) Peak rank
E) Weeks on board
Note: - From the home page you have to click on the charts option then hot 100-page link through code.

In [34]:
```
# connecting to the web driver
driver=webdriver.Chrome(r"C:/Users/HP/Downloads/chromedriver_win32 (1)/chromedriver.exe")

# getting the webpage of mentioned url
url = ("https://www.billboard.com/")
driver.get(url)
```

In [36]:
```
# clicking on option button
charts=driver.find_element_by_xpath("//a[@class='header__main-link header__main-link--charts']").click()
```

In [38]:
```
# creating empty lists
Song_Name = []
Artist_Name =[]
Last_week_rank = []
Peak_rank = []
Weeks_on_board = []

# getting urls for top 100 songs
urls = driver.find_element_by_xpath("//li[@class='header__submenu__list__element']//a")
page_url = urls.get_attribute("href")
driver.get(page_url)
time.sleep(4)

# scraping data of song names
for i in driver.find_elements_by_xpath("//span[@class='chart-element__information__song text--truncate color--primary']"):
    Song_Name.append(i.text)

# scraping data of artist names
for i in driver.find_elements_by_xpath("//span[@class='chart-element__information__artist text--truncate color--secondary']"):
    Artist_Name.append(i.text)

# scraping data of last week ranks
for i in driver.find_elements_by_xpath("//div[@class='chart-element__meta text--center color--secondary text--last']"):
    Last_week_rank.append(i.text)


# scraping data of peak ranks
for i in driver.find_elements_by_xpath("//div[@class='chart-element__meta text--center color--secondary text--peak']"):
    Peak_rank.append(i.text)


# scraping data of weeks on board
for i in driver.find_elements_by_xpath("//div[@class='chart-element__meta text--center color--secondary text--week']"):
    Weeks_on_board.append(i.text)


# creating dataframe for scraped data
billiboard = pd.DataFrame({})
billiboard['Name'] = Song_Name
billiboard['Artist'] = Artist_Name
billiboard['Last Week Rank'] = Last_week_rank
billiboard['Peak Rank'] = Peak_rank
billiboard['Weeks on board'] = Weeks_on_board
billiboard
```

Out[38]:

	Name	Artist	Last Week Rank	Peak Rank	Weeks on board
0	Easy On Me	Adele	1	1	3
1	Stay	The Kid LAROI & Justin Bieber	2	1	16
2	Industry Baby	Lil Nas X & Jack Harlow	3	1	14
3	Fancy Like	Walker Hayes	4	3	19
4	Bad Habits	Ed Sheeran	5	2	18
...	...	...	...	...	...
95	To Be Loved By You	Parker McCollum	-	96	1
96	Ain't Shit	Doja Cat	-	24	14
97	Life Goes On	Oliver Tree	-	91	2
98	Come Through	H.E.R. Featuring Chris Brown	-	64	16
99	Nevada	YoungBoy Never Broke Again	-	58	4

100 rows × 5 columns

In [39]:
```
driver.close()
```

In [ ]:

In [ ]:

Q 7 : Scrape the details of Data science recruiters from naukri.com.

Url = https://www.naukri.com/
You have to find the following details:
A) Name
B) Designation
C) Company
D) Skills they hire for
E) Location
Note: - From naukri.com homepage click on the recruiters option and the on the search pane type Data science and click on search. All this should be done through code

In [2]:
```
# connecting to the web driver
driver=webdriver.Chrome(r"C:/Users/HP/Downloads/chromedriver_win32 (1)/chromedriver.exe")

# getting the webpage of mentioned url
url = ("https://www.naukri.com/")
driver.get(url)
time.sleep(3)
```

In [3]:
```
# fetching urls to navigate recruiter page
recruiter = driver.find_element_by_xpath("//a[@title='Search Recruiters']")
page_url = recruiter.get_attribute("href")

driver.get(page_url)
time.sleep(3)
```

In [4]:
```
# fetching search button, sending keys and clicking on it
search = driver.find_element_by_xpath("//div[@class='inpWrap']//input")
search.send_keys("Data Science")

btn = driver.find_element_by_xpath("//button[@class='fl qsbSrch blueBtn']").click()
time.sleep(3)
```

In [14]:
```
# creating empty lists
Name = []
Designation = []
Company = []
Skills = []
Location = []

# scraping data of Names
for i in driver.find_elements_by_xpath("//span[@class='fl ellipsis']"):
    Name.append(i.text)
time.sleep(3)


# scraping data of Designation
for i in driver.find_elements_by_xpath("//span[@class='ellipsis clr']"):
    Designation.append(i.text)
time.sleep(3)


# scraping data of Company Name
for i in driver.find_elements_by_xpath("//div[@class='vcard']//p[1]/a[2]"):
    Company.append(i.text)
time.sleep(3)


# scraping data of Skills
for i in driver.find_elements_by_xpath("//div[@class='hireSec highlightable']"):
    try:
        if i.text == "Not Specified": raise NoSuchElementException
        Skills.append(i.text)
    except NoSuchElementException:
        Skills.append('-')
time.sleep(3)


# scraping data of Location
```

```
for i in driver.find_elements_by_xpath("//div[@class='vcard']//p[1]/span/small"):
    try:
        if i.text == "Not Specified": raise NoSuchElementException
        Location.append(i.text)
    except NoSuchElementException:
        Location.append('-')
time.sleep(3)


# creating dataframe for scraped data
Naukri = pd.DataFrame({})
Naukri['Name'] = Name[:49]
Naukri['Designation'] = Designation[:49]
Naukri['Company'] = Company[:49]
Naukri['Skills'] = Skills[:49]
Naukri['Location'] = Location[:49]
Naukri
```

Out[14]:

	Name	Designation	Company	Skills	Location
0	Aakash Harit	HR Manager	Data Science Network	Classic ASP Developer, Internet Marketing Prof...	Delhi
1	shravan Kumar Gaddam	Company Recruiter	Shore Infotech India Pvt. Ltd	.Net, Java, Data Science, Linux Administration...	Hyderabad / Secunderabad
2	MARSIAN Technologies LLP	Company HR	MARSIAN Technologies LLP	Data Science, Artificial Intelligence, Machine...	Pune
3	Anik Agrawal	Company Recruiter	Enerlytics Software Solutions Pvt Ltd	Mean Stack, javascript, angularjs, mongodb, We...	Ahmedabad
4	subhas patel	Founder CEO	LibraryXProject	Hadoop, Spark, Digital Strategy, Data Architec...	UK - (london)
5	Abhishek - Only Analytics Hiring - India and	Recruitment Lead Consultant	Apidel Technologies Division of Transpower	Analytics, Business Intelligence, Business Ana...	Vadodara / Baroda
6	Institute for Financial Management and Resear	Programme Manager	IFMR	Data Science	Chennai
7	Narasimha	Company Recruiter	Step Next Private Limited	Data Modeling, Data Wrangling, seaborn, eda, p...	Hyderabad / Secunderabad
8	Balu Ramesh	HR Administrator	Techvantage Systems Pvt Ltd	Machine Learning, algorithms, Go Getter, Compu...	Trivandrum
9	Asif Lucknowi	Director	Weupskill- Live Wire India	Technical Training, Software Development, Pres...	Indore
10	InstaFinancials	Human Resource	CBL Data Science Private Limited	Software Development, It Sales, Account Manage...	Bengaluru / Bangalore
11	Priyanka Akiri	HR Manager	Infinitive Software Solutions	Oracle Dba, Data Science, Data Warehousing, ET...	Hyderabad
12	Kalpana Dumpala	Executive Hiring	Innominds Software	Qa, Ui/ux, Java Developer, Java Architect, C++...	Hyderabad / Secunderabad
13	Mubarak	Company HR	MoneyTap	Business Intelligence, Data Warehousing, Data ...	Bengaluru / Bangalore
14	Kushal Rastogi	Company HR	QuantMagnum Technologies Pvt. Ltd.	Office Administration, Hr Administration, tele...	Mumbai
15	Ruchi Dhote	Senior Executive Talent Acquisition	Bristlecone India Ltd	Qlikview, Qlik Sense, Microsoft Azure, Power B...	Pune
16	Kapil Devang	HR Manager	BISP Solutions	Big Data, Hadoop, Data Analytics, Data Science	Bhopal
17	Mahesh Babu Channa	HR Team Lead	SocialPrachar.com	Social Media, digital media maketing, seo, smm...	Hyderabad / Secunderabad
18	Sandhya Khandagale	HR Recruiter	Compumatrice Multimedia Pvt Ltd	Big Data, Data Science, Artificial Intelligenc...	Pune
19	Manisha Yadav	HR Executive	Easi Tax	Telecalling, Client Interaction, Marketing, Re...	Navi Mumbai
20	Riya Rajesh	Manager Talent Acquisition	Novelworx Digital Solutions	Data Science	Cochin
21	Rashmi Bhattacharjee	HR Head	AXESTRACK SOFTWARE SOLUTIONS PRIVATE...	Corporate Sales, Software Development, Softwar...	Delhi
22	Faizan Kareem	HR MANAGER	FirstTech Consaltants Pvt.Ltd	Data Analytics, Data Science, Machine Learning...	Hyderabad / Secunderabad
23	Rithika dadwal	HR Recruiter	Affine Analytics	Data Science, Machine Learning, Python, R, Dee...	Pune
24	Shaun Rao	Manager Human Resources	Exela Technologies	Java, Net, Angularjs, Hr, Infrastructure, Mana...	Pune
25	Azahar Shaikh	Company Recruiter	NEAL ANALYTICS SERVICES PVT LTD	Data Science, Artificial Intelligence, Machine...	Pune
26	Manas	Lead Talent acquisition	Autumn Leaf Consulting Services Private...	Software Architecture, Vp Engineering, Product...	Bengaluru / Bangalore
27	kumar	Proprietor	trainin	Data Science, Hadoop, Rpas, Devops, Python, Aw...	Bengaluru / Bangalore
28	Sunil Vedula	CEO	Nanoprecise Sci Corp	Signal Processing, Machine Learning, Neural Ne...	Delhi
29	Rajat Kumar	Founder CEO	R.S Consultancy & Services	Web Technologies, Project Management, Software...	Mysoru / Mysore
30	Jayanth N	Project Manager	Dollarbird Information Services Pvt, Ltd	Data Analytics, Managed Services, Team Leading...	Bengaluru / Bangalore
31	Priya Khare	Senior Manager	Independent Consultant	Data Science, Artificial Intelligence, analyti...	Bengaluru / Bangalore
32	Dhruv Dev Dubey	Company Recruitment Head	Confidential	Server Administartion, Verilog, Vhdl, Digital ...	Hyderabad / Secunderabad
33	SREEDHAR	Recruitment Consultant	JOBSMILL BUSINESS SOLUTIONS PRIVATE LIMITED	Data Science, Machine Learning, Big Data Analy...	Ghaziabad
34	Ravi Dubey	Recruitment Manager	HyrEzy Talent Solutions LLP	Walmart Interra Skeps Expressstores indifi whi...	Bengaluru / Bangalore
35	Radha Manivasagam	HR Executive	Techcovery	Python, Artificial Intelligence, Machine Learn...	Noida
36	Prateek Kumar	Head	Trisect	Java, Python, Angularjs, Software Testing, Mac...	New Delhi
37	Amit Sharma	Consultant	ASCO consulting	Machine Learning, Artificial Intelligence, Dat...	Chennai

38	Kanan	senior technology instructor	NY INST	C, C++, Artificial Intelligence, Python, Php, ...	Aligarh
39	Shashikant Chaudhary	HR Recruiter/HR Excutive	3D India Staffing Research & Consulting...	Relationship Management, Retail Sales, Private...	Salt Lake City
40	Brad	Manager, Technical Recruiting	O.C. Tanner	Data Science, Software Engineering	Pune
41	Rutuja Pawar	Technical Recruiter	Demand Matrix	Data Science, Big Data Analytics, Digital Mark...	Bengaluru / Bangalore
42	Madhusudhan Sridhar	Erp Implementer	MADHUSUDHAN SRIDHAR	Data Science, Recruitment, Salary	Mumbai
43	Ankit Sinha	Head Analytics	Suntech Global	B.Tech, Tableau, Statistics, R, Analytics, Tim...	Indore
44	Gaurav Chouhan	Chief Technical Officer	Strategic Consulting Lab	Software Development, Business Intelligence, B...	Bengaluru / Bangalore
45	Rashi Kacker	Sr Product Manager	Impel Labs Pvt. Ltd.	Data Science, Node.js, Angularjs	MYSORE
46	Ashwini	Director Global Delivery	MRP Advisers	Data Science, Media Marketing, Resource Planni...	Hyderabad / Secunderabad
47	Balaji Kolli	Co Founder	Saras Solutions India Pvt Ltd	Data Analysis, Learning, Data Science, Compute...	Bengaluru / Bangalore
48	Rajani Nagaraj	HR Manager	WildJasmine	Java, Hadoop, R, Machine Learning, Spark, Flum...	Mumbai

In [17]:
```
driver.close()
```

In [ ]:

In [ ]:

Q 8 : 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/"
You have to find the following details:
A) Book name
B) Author name
C) Volumes sold
D) Publisher
E) Genre

In [18]:
```
# connecting to the web driver
driver=webdriver.Chrome(r"C:/Users/HP/Downloads/chromedriver_win32 (1)/chromedriver.exe")

# getting the webpage of mentioned url
url = ("https://www.theguardian.com/news/datablog/2012/aug/09/best-selling-books-all-time-fifty-shades-grey-compare/")
driver.get(url)
time.sleep(3)
```

In [19]:
```
# creating empty lists
Book_name = []
Author_name = []
Volumes_sold = []
Publisher = []
Genre = []


# scraping book names data
for i in driver.find_elements_by_xpath("//tbody//tr//td[2]"):
    Book_name.append(i.text)


# scraping author names data
for i in driver.find_elements_by_xpath("//tbody//tr//td[3]"):
    try:
        if i.text == '0' : raise NoSuchElementException
        Author_name.append(i.text)
    except NoSuchElementException:
        Author_name.append('-')
time.sleep(3)


# scraping data of volumes sold
for i in driver.find_elements_by_xpath("//tbody//tr//td[4]"):
    Volumes_sold.append(i.text)


# scraping data of publisher names
for i in driver.find_elements_by_xpath("//tbody//tr//td[5]"):
    Publisher.append(i.text)


# scraping  data of genre
for i in driver.find_elements_by_xpath("//tbody//tr//td[6]"):
    Genre.append(i.text)


# creating dataframe for scraped data
Novels = pd.DataFrame({})
Novels['Book Name'] = Book_name
Novels['Author'] = Author_name
Novels['Volume sold'] = Volumes_sold
Novels['Publisher'] = Publisher
Novels['Genre'] = Genre
Novels
```

Out[19]:

	Book Name	Author	Volume 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 [20]:
```
driver.close()
```

In [ ]:

In [ ]:

Q 9 : Scrape the details most watched tv series of all time from imdb.com.

Url = https://www.imdb.com/list/ls095964455/
You have to find the following details:
A) Name
B) Year span
C) Genre
D) Run time
E) Ratings
F) Votes

In [21]:
```
# connecting to the web driver
driver=webdriver.Chrome(r"C:/Users/HP/Downloads/chromedriver_win32 (1)/chromedriver.exe")

# getting the webpage of mentioned url
url = ("https://www.imdb.com/list/ls095964455/")
driver.get(url)
```

In [22]:
```
# creating empty lists
Name = []
Year_span = []
Genre = []
Run_time = []
Ratings = []
Votes = []

# scraped data of Names
for i in driver.find_elements_by_xpath("//h3[@class='lister-item-header']/a"):
    Name.append(i.text)


# scraped data of Year span
for i in driver.find_elements_by_xpath("//span[@class='lister-item-year text-muted unbold']"):
    Year_span.append(i.text)


# scraped data of Genre
for i in driver.find_elements_by_xpath("//span[@class='genre']"):
    Genre.append(i.text)


# scraped data of Run time
for i in driver.find_elements_by_xpath("//span[@class='runtime']"):
    Run_time.append(i.text)


# scraped data of Ratings
for i in driver.find_elements_by_xpath("//div[@class='ipl-rating-star small']//span[2]"):
    Ratings.append(i.text)


# scraped data of Votes
for i in driver.find_elements_by_xpath("//div[@class='lister-item-content']//p[4]/span[2]"):
    Votes.append(i.text)


# creating dataframe for scraped data
TV_Series = pd.DataFrame({})
TV_Series['Name'] = Name
TV_Series['Year Span'] = Year_span
TV_Series['Genre'] = Genre
TV_Series['Run Time'] = Run_time
TV_Series['Ratings'] = Ratings
TV_Series['Votes'] = Votes
TV_Series
```

Out[22]:

	Name	Year Span	Genre	Run Time	Ratings	Votes
0	Game of Thrones	(2011–2019)	Action, Adventure, Drama	57 min	9.2	1,894,576
1	Stranger Things	(2016–)	Drama, Fantasy, Horror	51 min	8.7	925,873
2	The Walking Dead	(2010–2022)	Drama, Horror, Thriller	44 min	8.2	910,239
3	13 Reasons Why	(2017–2020)	Drama, Mystery, Thriller	60 min	7.5	272,388
4	The 100	(2014–2020)	Drama, Mystery, Sci-Fi	43 min	7.6	232,775

...	...	...	...	...	...	...
95	Reign	(2013–2017)	Drama, Fantasy	42 min	7.5	46,260
96	A Series of Unfortunate Events	(2017–2019)	Adventure, Comedy, Drama	50 min	7.8	57,102
97	Criminal Minds	(2005–2020)	Crime, Drama, Mystery	42 min	8.1	179,246
98	Scream: The TV Series	(2015–2019)	Comedy, Crime, Drama	45 min	7.1	37,535
99	The Haunting of Hill House	(2018)	Drama, Horror, Mystery	572 min	8.6	210,665

100 rows × 6 columns

In [23]:
```
driver.close()
```

In [ ]:

In [ ]:

Q 10 : Details of Datasets from UCI machine learning repositories.

```
Url = https://archive.ics.uci.edu/
You have to find the following details:
A) Dataset name
B) Data type
C) Task
D) Attribute type
E) No of instances
F) No of attribute
G) Year
 Note: - from the home page you have to go to the Show All Dataset page through code.
```

In [7]:
```
# connecting to the web driver
driver=webdriver.Chrome(r"C:/Users/HP/Downloads/chromedriver_win32 (1)/chromedriver.exe")

# getting the webpage of mentioned url
url = (" https://archive.ics.uci.edu/")
driver.get(url)
```

In [8]:
```
# fetching view all dataset button from the webpage
viewall_dataset = driver.find_element_by_xpath("//tbody[1]//tr/td[2]/span[2]/a")
page_url = viewall_dataset.get_attribute("href")
driver.get(page_url)
time.sleep(3)
```

In [9]:
```
# fetching page urls of all datasets
view_list = driver.find_element_by_xpath("/html/body/table[2]/tbody/tr/td[2]/table[1]/tbody/tr/td[2]/p/a")
list_url = view_list.get_attribute("href")
driver.get(list_url)
time.sleep(3)
```

In [10]:
```
# fetching urls for each dataset
dataset_url = driver.find_elements_by_xpath("//p[@class='normal']//b/a")

urls = []
for i in dataset_url:
    urls.append(i.get_attribute("href"))
```

In [ ]:
```
# creating empty lists
Dataset_name = []
Data_type = []
Task = []
Attribute_type = []
No_of_instances = []
No_of_attributes = []
Year = []

for i in urls:
    driver.get(i)
    time.sleep(3)


    # scraping  Dataset name
    try:
        dataset_name = driver.find_element_by_xpath("//span[@class='heading']")
        Dataset_name.append(dataset_name.text)
    except NoSuchElementException:
        Dataset_name.append('-')
    time.sleep(3)



    # scraping data type
    try:
        data_type = driver.find_element_by_xpath("//table[@border='1']//tbody/tr/td[2]")
        if data_type.text == "N/A": raise NoSuchElementException
        Data_type.append(data_type.text)
    except NoSuchElementException:
        Data_type.append('-')
    time.sleep(3)


    # scraping Task
    try:
        task = driver.find_element_by_xpath("//table[@border='1']//tbody/tr[3]/td[2]")
        if task.text == "N/A": raise NoSuchElementException
        Task.append(task.text)
    except NoSuchElementException:
```

```
        Task.append('-')
    time.sleep(3)



    # scraping Attribute type
    try:
        attribute_type = driver.find_element_by_xpath("//table[@border='1']//tbody/tr[2]/td[2]")
        if attribute_type.text == "N/A": raise NoSuchElementException
        Attribute_type.append(attribute_type.text)
    except NoSuchElementException:
        Attribute_type.append('-')
    time.sleep(3)



    # scraping No of Instances
    try:
        instances = driver.find_element_by_xpath("//table[@border='1']//tbody/tr/td[4]")
        if instances.text == "N/A": raise NoSuchElementException
        No_of_instances.append(instances.text)
    except NoSuchElementException:
        No_of_instances.append('-')
    time.sleep(3)



    # scraping No of Arrtibutes
    try:
        attribute = driver.find_element_by_xpath("//table[@border='1']//tbody/tr[2]/td[4]")
        if attribute.text == "N/A": raise NoSuchElementException
        No_of_attributes.append(attribute.text)
    except NoSuchElementException:
        No_of_attributes.append('-')
    time.sleep(3)


    # scraping Year
    try:
        year = driver.find_element_by_xpath("//table[@border='1']//tbody/tr[2]/td[6]")
        if year.text == "N/A": raise NoSuchElementException
        Year.append(year.text[:4])
    except NoSuchElementException:
        Year.append('-')
    time.sleep(3)
```

In [ ]:

```
# creating dataframe for scraped data
ML = pd.DataFrame({})
ML['Data Name'] = Data_name
ML['Data Type '] = Data_type
ML['Task '] = Task
ML['Attribute Type '] = Attribute_type
ML['No of Instance '] = No_of_instances
ML['No of Attributes '] = No_of_attributes
ML['Year '] = Year
ML
```

In [ ]:

```
driver.close()
```

In [ ]:

In [ ]:

In [ ]:

In [ ]: