In [8]:
```
import re


import selenium
import pandas as pd
from selenium import webdriver

from bs4 import BeautifulSoup
import requests


import time


from selenium.common.exceptions import NoSuchElementException, StaleElementReferen

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

In [2]:
```
driver=webdriver.Chrome()
```

In [3]:
```
url = 'https://en.wikipedia.org/wiki/List_of_most-viewed_YouTube_videos'
driver.get(url)
```

In [4]:
```
Rank = []
Name = []
Artist = []
Date = []
Views = []
```

In [6]:
```
try:
    for i in driver.find_elements(By.XPATH,"//table[@class='wikitable sortable jque
        Rank.append(i.text)
except NoSuchElementException:
    Rank.append("-")


try:
    for i in driver.find_elements(By.XPATH,"//table[@class='wikitable sortable jque
        Name.append(i.text)
except NoSuchElementException:
    Name.append("-")

try:
    for i in driver.find_elements(By.XPATH,"//table[@class='wikitable sortable jque
        Artist.append(i.text)
except NoSuchElementException:
    Artist.append("-")


try:
    for i in driver.find_elements(By.XPATH,"//table[@class='wikitable sortable jque
        Date.append(i.text)
except NoSuchElementException:
    Date.append("-")


try:
    for i in driver.find_elements(By.XPATH,"//table[@class='wikitable sortable jque
        Views.append(i.text)
except NoSuchElementException:
    Views.append("-")
```

```
Wiki = pd.DataFrame({})
Wiki['Rank'] = Rank
Wiki['Name'] = Name
Wiki['Artist'] = Artist
Wiki['Upload Date'] = Date
Wiki['Views (in Billions)'] = Views


Wiki.Name = Wiki.Name.apply(lambda x:x[:-4].strip('"'))
Wiki
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Input In [6], in <cell line: 1>()
      1 try:
----> 2     for i in driver.find_elements(By.XPATH,"//table[@class='wikitable sort
able jquery-tablesorter'][1]/tbody/tr/td[1]"):
      3         Rank.append(i.text)
      4 except NoSuchElementException:

NameError: name 'By' is not defined
```

In [ ]:
```
#question2:Scrape the details team India's international fixtures from bcci.tv
```

In [9]:
```
driver=webdriver.Chrome()
```

In [10]:
```
url=('https://www.bcci.tv/')
driver.get(url)
```

In [13]:
```
btn=driver.find_elements(By.XPATH,"//div[@class='navigation__drop-down drop-down d
driver.get(btn.get_attribute("href"))
time.sleep(3)

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

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Input In [13], in <cell line: 1>()
----> 1 btn=driver.find_elements(By.XPATH,"//div[@class='navigation__drop-down dro
p-down drop-down--reveal-on-hover']/div/ul/li/a")
      2 driver.get(btn.get_attribute("href"))
      3 time.sleep(3)

NameError: name 'By' is not defined
```

In [ ]:
```

```

In [ ]:
```

```

In [14]:
```
for i in driver.find_elements(By.XPATH,("//div[@class='fixture__format-strip']/spa
    Match_Title.append(i.text)

for i in driver.find_elements(By.XPATH,("//div[@class='fixture__format-strip']/spa
    Series.append(i.text)
```

```
for i in driver.find_elements(By.XPATH,"//div[@class='fixture__description u-unskew
    Place.append(i.text)

for i in driver.find_elements(By.XPATH,"//span[@class='fixture__datetime tablet-on
    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
```

```
  Input In [14]
    for i in driver.find_elements(By.XPATH,("//div[@class='fixture__format-stri
p']/span[@class='u-unskewed-text fixture__format']"):

^
SyntaxError: invalid syntax
```

In [15]:
```
#Question3:
```

In [16]:
```
driver=webdriver.Chrome()
```

In [18]:
```
url = ("https://statisticstimes.com/")
driver.get(url)
```

In [ ]:
```
```

In [20]:
```
# clicking on Economy button
driver.find_elements(By.XPATH,"//div[@class='navbar']/div[2]/button").click()

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

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

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Input In [20], in <cell line: 2>()
      1 # clicking on Economy button
----> 2 driver.find_elements(By.XPATH,"//div[@class='navbar']/div[2]/button").clic
k()
      4 # clicking on India
      5 driver.find_elements(By.XPATH,"//div[@class='dropdown-content']/a[3]").cli
ck()

NameError: name 'By' is not defined
```

In [23]:
```
Rank = []
State = []
GSDP1 = []
GSDP2 = []
```

```
Share = []
GDP_billion = []
# scraping Rank
try:
    for i in driver.find_elements(By.XPATH,"//table[@class='display dataTable']/tb
        Rank.append(i.text)
except NoSuchElementException:
    Rank.append("_")

# scraping State
try:
    for i in driver.find_elements(By.XPATH,"//table[@class='display dataTable']/tb
        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']/tb
        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']/tb
        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']/tb
        Share.append(i.text)
except NoSuchElementException:
    Share.append("_")

# scraping GDP $ billion
try:
    for i in driver.find_elements(By.XPATH,"//table[@class='display dataTable']/tb
        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
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Input In [23], in <cell line: 8>()
      7 # scraping Rank
      8 try:
----> 9     for i in driver.find_elements(By.XPATH,"//table[@class='display dataTa
ble']/tbody/tr/td[1]"):
     10         Rank.append(i.text)
     11 except NoSuchElementException:

NameError: name 'By' is not defined
```

In [ ]:
```
#question:
```

In [24]:
```
driver=webdriver.Chrome()
```

In [25]:
```
url = ("https://github.com/")
driver.get(url)
```

In [26]:
```
# getting explore button and clicking on it
explore = driver.find_elements(By.XPATH,"/html/body/div[1]/header/div/div[2]/nav/ul

# selecting trending option
trend_url = driver.find_elements(By.XPATH,"/html/body/div[1]/header/div/div[2]/nav/
urls = trend_url.get_attribute("href")
driver.get(urls)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Input In [26], in <cell line: 2>()
      1 # getting explore button and clicking on it
----> 2 explore = driver.find_elements(By.XPATH,"/html/body/div[1]/header/div/div
[2]/nav/ul/li[4]/details").click()
      4 # selecting trending option
      5 trend_url = driver.find_elements(By.XPATH,"/html/body/div[1]/header/div/di
v[2]/nav/ul/li[4]/details/div/ul[2]/li[3]/a")

NameError: name 'By' is not defined
```

In [28]:
```
# creating empty list
URLs = []
repository_title = []
Description = []
Contributors = []
Language = []
lang = []


repository = driver.find_elements(By.XPATH,"//h1[@class='h3 lh-condensed']//a")
for i in repository:
    URLs.append(i.get_attribute("href"))


title = driver.find_elements(By.XPATH,"//h1[@class = 'h3 lh-condensed']")
for i in title:
    repository_title.append(i.text)


for i in URLs:
    driver.get(i)
    time.sleep(5)
```

```
    try:
        desc = driver.ffind_elements(By.XPATH,"//p[@class='f4 mt-3']")
        Description.append(desc.text)
    except NoSuchElementException:
        Description.append('-')


    try:
        contributor = driver.find_elements(By.XPATH,"//*[contains(text(),'    Contr
        Contributors.append(contributor.text.replace('Contributors',''))
    except NoSuchElementException:
        Contributors.append('-')



    try:
        for i in driver.find_elements(By.XPATH,"//ul[@class= 'list-style-none']//li
            lang.append(i.text)
        Language.append(lang)
    except NoSuchElementException:
        Language.append('-')


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

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Input In [28], in <cell line: 10>()
      6 Language = []
      7 lang = []
---> 10 repository = driver.find_elements(By.XPATH,"//h1[@class='h3 lh-condense
d']//a")
     11 for i in repository:
     12     URLs.append(i.get_attribute("href"))

NameError: name 'By' is not defined
```

In [ ]:
```
#question
```

In [29]:
```
driver=webdriver.Chrome()
```

In [30]:
```
url = ("https://www.billboard.com/")
driver.get(url)
```

In [ ]:
```
charts=driver.find_elements(By.XPATH,"//a[@class='header__main-link header__main-li
```

In [ ]:
```

```

In [32]:
```
Song_Name = []
Artist_Name =[]
Last_week_rank = []
Peak_rank = []
Weeks_on_board = []


urls = driver.find_elements(By.XPATH,"//li[@class='header__submenu__list__element'
```

```
page_url = urls.get_attribute("href")
driver.get(page_url)
time.sleep(4)


for i in driver.find_elements(By.XPATH,"//span[@class='chart-element__information__
    Song_Name.append(i.text)

for i in driver.find_elements(By.XPATH,"//span[@class='chart-element__information__
    Artist_Name.append(i.text)


for i in driver.find_elements(By.XPATH,"//div[@class='chart-element__meta text--cen
    Last_week_rank.append(i.text)



for i in driver.find_elements(By.XPATH,"//div[@class='chart-element__meta text--cen
    Peak_rank.append(i.text)



for i in driver.find_elements(By.XPATH,"//div[@class='chart-element__meta text--cen
    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
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Input In [32], in <cell line: 8>()
      4 Peak_rank = []
      5 Weeks_on_board = []
----> 8 urls = driver.find_elements(By.XPATH,"//li[@class='header__submenu__list__
element']//a")
      9 page_url = urls.get_attribute("href")
     10 driver.get(page_url)

NameError: name 'By' is not defined
```

In [ ]:
```
#question
```

In [33]:
```
driver=webdriver.Chrome()
```

In [34]:
```
url = (" https://archive.ics.uci.edu/")
driver.get(url)
```

In [37]:
```
viewall_dataset = driver.find_elements(By.XPATH,"//tbody[1]//tr/td[2]/span[2]/a")
page_url = viewall_dataset.get_attribute("href")
driver.get(page_url)
time.sleep(3)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Input In [37], in <cell line: 1>()
----> 1 viewall_dataset = driver.find_elements(By.XPATH,"//tbody[1]//tr/td[2]/span
[2]/a")
      2 page_url = viewall_dataset.get_attribute("href")
      3 driver.get(page_url)

NameError: name 'By' is not defined
```

In [38]:
```
view_list = driver.find_elements(By.XPATH,"/html/body/table[2]/tbody/tr/td[2]/table
list_url = view_list.get_attribute("href")
driver.get(list_url)
time.sleep(3)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Input In [38], in <cell line: 1>()
----> 1 view_list = driver.find_elements(By.XPATH,"/html/body/table[2]/tbody/tr/td
[2]/table[1]/tbody/tr/td[2]/p/a")
      2 list_url = view_list.get_attribute("href")
      3 driver.get(list_url)

NameError: name 'By' is not defined
```

In [39]:
```
dataset_url = driver.find_elements(By.XPATH,"//p[@class='normal']//b/a")

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

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Input In [39], in <cell line: 1>()
----> 1 dataset_url = driver.find_elements(By.XPATH,"//p[@class='normal']//b/a")
      3 urls = []
      4 for i in dataset_url:

NameError: name 'By' is not defined
```

In [40]:
```
Dataset_name = []
Data_type = []
Task = []
Attribute_type = []
No_of_instances = []
No_of_attributes = []
Year = []
for i in urls:
    driver.get(i)
    time.sleep(3)
    try:
        dataset_name = driver.find_elements(By.XPATH,"//span[@class='heading']")
        Dataset_name.append(dataset_name.text)
    except NoSuchElementException:
        Dataset_name.append('-')
    time.sleep(3)
    try:
        data_type = driver.find_elements(By.XPATH,"//table[@border='1']//tbody/tr/t
        if data_type.text == "N/A": raise NoSuchElementException
        Data_type.append(data_type.text)
    except NoSuchElementException:
        Data_type.append('-')
    time.sleep(3)
    try:
```

```
        task = driver.find_elements(By.XPATH,"//table[@border='1']//tbody/tr[3]/td
        if task.text == "N/A": raise NoSuchElementException
        Task.append(task.text)
    except NoSuchElementException:
        Task.append('-')
    time.sleep(3)
     try:
        attribute_type = driver.find_elements(By.XPATH,"//table[@border='1']//tbody
        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_elements(By.XPATH,"//table[@border='1']//tbody/tr/
        if instances.text == "N/A": raise NoSuchElementException
        No_of_instances.append(instances.text)
    except NoSuchElementException:
        No_of_instances.append('-')
    time.sleep(3)



    try:
        attribute = driver.find_elements(By.XPATH,"//table[@border='1']//tbody/tr[
        if attribute.text == "N/A": raise NoSuchElementException
        No_of_attributes.append(attribute.text)
    except NoSuchElementException:
        No_of_attributes.append('-')
    time.sleep(3)

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

```
  Input In [40]
    try:
    ^
IndentationError: unexpected indent
```

In [41]:
```
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
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Input In [41], in <cell line: 2>()
      1 ML = pd.DataFrame({})
----> 2 ML['Data Name'] = Data_name
      3 ML['Data Type '] = Data_type
      4 ML['Task '] = Task

NameError: name 'Data_name' is not defined
```

In [ ]:
```
#question:
```

In [42]:
```
driver=webdriver.Chrome()
```

In [43]:
```
url = ("https://www.imdb.com/list/ls095964455/")
driver.get(url)
```

In [45]:
```
# 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
    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']//spa
    Ratings.append(i.text)


# scraped data of Votes
for i in driver.find_elements(By.XPATH,"//div[@class='lister-item-content']//p[4]/
    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
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Input In [45], in <cell line: 10>()
      7 Votes = []
      9 # scraped data of Names
---> 10 for i in driver.find_elements(By.XPATH,"//h3[@class='lister-item-header']/
a"):
     11     Name.append(i.text)
     14 # scraped data of Year span

NameError: name 'By' is not defined
```

In [ ]:
```
#questtion:
```

In [46]:
```
driver=webdriver.Chrome()
```

In [47]:
```
url = ("https://www.theguardian.com/news/datablog/2012/aug/09/best-selling-books-a
driver.get(url)
time.sleep(3)
```

In [48]:
```
Book_name = []
Author_name = []
Volumes_sold = []
Publisher = []
Genre = []



for i in driver.find_elements(By.XPATH,"//tbody//tr//td[2]"):
    Book_name.append(i.text)



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)


for i in driver.find_elements(By.XPATH,"//tbody//tr//td[4]"):
    Volumes_sold.append(i.text)


for i in driver.find_elements(By.XPATH,"//tbody//tr//td[5]"):
    Publisher.append(i.text)



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
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Input In [48], in <cell line: 9>()
      4 Publisher = []
      5 Genre = []
----> 9 for i in driver.find_elements(By.XPATH,"//tbody//tr//td[2]"):
     10     Book_name.append(i.text)
     14 for i in driver.find_elements(By.XPATH,"//tbody//tr//td[3]"):

NameError: name 'By' is not defined
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: