WEB SCRAPING USING SELENIUM AND BEAUTIFULSOUP

In [1]:

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

In [2]:

```python
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.common.exceptions import TimeoutException
```

Extract page using selenium

In [3]:

```python
driver = webdriver.Chrome(executable_path='\chromeDriver\chromedriver.exe')
driver.get('https://www.konga.com/category/accessories-computing-5227')
```

Passing the extracted page to Beautifulsoup

In [4]:

```python
page=bs(driver.page_source)
```

In [5]:

```python
page
```

Out[5]:

```
<!DOCTYPE html>
<html lang="en" prefix="og: http://ogp.me/ns#" (http://ogp.me/ns#") style
="position: static;" xmlns="http://www.w3.org/1999/xhtml"><head><meta char
set="utf-8"/><meta content="ie=edge" http-equiv="x-ua-compatible"/><meta c
ontent="initial-scale=1, width=device-width" name="viewport"/><meta conten
t="#ed017f" name="theme-color"/><meta content="#ed017f" name="msapplicatio
n-TileColor"/><link href="/manifest.json" rel="manifest"/><meta content="#
ed017f" name="msapplication-TileColor"/><meta content="yes" name="mobile-w
eb-app-capable"/><meta content="Konga" name="application-name"/><link href
="/static/android-chrome-192x192.png" rel="icon" sizes="192x192"/><meta co
ntent="yes" name="apple-mobile-web-app-capable"/><meta content="#ed017f" n
ame="apple-mobile-web-app-status-bar-style"/><meta content="Konga" name="a
pple-mobile-web-app-title"/><link href="/static/apple-touch-icon-150x150.p
ng" rel="apple-touch-icon"/><meta content="summary_large_image" name="twit
ter:card"/><meta content="@ShopKonga" name="twitter:site"/><meta content
="website" property="og:type"/><meta content="en-US" property="og:locale"/
><meta content="image/png" property="og:image:type"/><meta content="400" p
```

In [6]:

```
products_section= page.find("ul", {"class" : "b49ee_2pjyI _58c31_2R34y _22339_3gQb9"})
```

In [7]:

```
product_info=products_section.find_all("li",{"class":"bbe45_3oExY _22339_3gQb9"})
```

In [8]:

```
# Number of item available on each page
len(product_info)
```

Out[8]:

20

Extract all the infornation needed from an instance of item

In [33]:

```
product_info[2].find("span",{"class":"_4472a_zYlL- _6c244_q2qap"}).text.split()[0]
```

Out[33]:

'27%'

In [34]:

```
product_info[2].find("div",{"class":"af885_1iPzH"}).text
```

Out[34]:

'Zinox Z100 Gtx Notebook - Intel Cherr...'

In [35]:

```
product_info[2].find("span",{"class":"d7c0f_sJAqi"}).text.strip()
```

Out[35]:

'₦59,500'

In [36]:

```
product_info[2].find("span",{"class":"_7cc7b_23GsY"}).text.split()[-1]
```

Out[36]:

'ashour-fzco'

In [17]:

```python
product_info[2].find("span",{"class":"eea9b_1Ma8-"}).text
```

Out[17]:

```
'(No reviews yet)'
```

In [18]:

```python
product_info[2].find("span",{"class":"_17030_3ZCyn"}).text
```

Create a function to extract data on one page

In [19]:

```python
def Get_data(product_info):
    store=[]
    products=product_info
    for info in products:
        infos={}
        try:
            infos['off %']=info.find("span",{"class":"_4472a_zYlL- _6c244_q2qap"}).text.spl
        except:
            pass
        infos['Description']=info.find("div",{"class":"af885_1iPzH"}).text
        infos['Price']=info.find("span",{"class":"d7c0f_sJAqi"}).text.strip()
        try:
            infos['number of review']=info.find("span",{"class":"eea9b_1Ma8-"}).text
            if infos['number of review']=='(No reviews yet)':
                infos['number of review']=0

        except:
            pass
        try:
            infos['Pick up']-info.find("div",{"class":"_09e22_1ojNd"}).text.split()[1:][0]
        except:
            pass
        try:
            infos['available at']=info.find("span",{"class":"_17030_3ZCyn"}).text
        except:
            pass
        store += [infos]
    return store
```

In [20]:

```python
data1=Get_data(product_info)
pd.DataFrame((data1))
```

Out[20]:

	off %	Description	Price	store name	number of review	available at
0	18%	HP Notebook 15-RB006NIA AMD A4 Dual C...	₦88,500	konga	4 Reviews	Lagos,Abuja,PH
1	19%	HP Notebook 15 Intel Celeron N3060 - ...	₦94,990	konga	1 Review	Lagos,Abuja,PH
2	27%	Zinox Z100 Gtx Notebook - Intel Cherr...	₦59,500	ashour-fzco	0	Lagos
3	NaN	HP Stream 11 - Intel X5-e8000 - 4GB R...	₦80,000	meeso	0	NaN
4	19%	HP 15 Laptop - Intel Core i3 - 1TB HD...	₦157,000	inc	0	NaN
5	49%	Unique Rechargeable Wireless Mouse - ...	₦3,500	ventures	0	Lagos

In [21]:

```python
# get the first and last page number from a page
page1=int(page.find("ul",{"class":"_6d32f_2_vdu"}).text[0])
last_page=int(page.find("ul",{"class":"_6d32f_2_vdu"}).text[-2:])
```

Data extraction of a particular section on all pages available

In [23]:

```python
links=[f"https://www.konga.com/category/accessories-computing-5227?page={i}" for i in range

def All_page_data(page_url):

    links=[f"{page_url}?page={i}" for i in range(page1,last_page+1,1)]
    data_store=[]
    for url in links:
        driver.get(url)
        page=bs(driver.page_source)
        products_section= page.find("ul", {"class" : "b49ee_2pjyI _58c31_2R34y _22339_3gQb9
        product_info=products_section.find_all("li",{"class":"bbe45_3oExY _22339_3gQb9"})
        data = get_data(product_info)
        data_store+=data
    return data_store
```

In [24]:

```python
page1_data=All_page_data('https://www.konga.com/category/accessories-computing-5227?page=1'
```

In [25]:

```python
pd.DataFrame(page1_data)
```

Out[25]:

	off %	Description	Price	store name	number of review	available at
0	18%	HP Notebook 15-RB006NIA AMD A4 Dual C...	₦88,500	konga	4 Reviews	Lagos,Abuja,PH
1	19%	HP Notebook 15 Intel Celeron N3060 - ...	₦94,990	konga	1 Review	Lagos,Abuja,PH
2	27%	Zinox Z100 Gtx Notebook - Intel Cherr...	₦59,500	ashour-fzco	0	Lagos
3	NaN	HP Stream 11 - Intel X5-e8000 - 4GB R...	₦80,000	meeso	0	NaN
4	19%	HP 15 Laptop - Intel Core i3 - 1TB HD...	₦157,000	inc	0	NaN
...	...	...	...	...	...	...
995	18%	HP Pavilion 15 Intel Core i5 - Touchs...	₦295,000	ojaonline2	0	NaN
996	NaN	HP 26a Laserjet Toner Cartridge	₦9,000	doublebigman	0	NaN
997	NaN	HP Pavilion 11 x360 Touch Screen - I...	₦168,000	ltd	0	NaN
998	19%	HP 15 Laptop - Intel Core i3 - 1TB HD...	₦159,000	gadgets	0	NaN
999	31%	Ergostand Laptop Notebook Stand And C...	₦3,100	synergy2020	0	NaN

1000 rows × 6 columns

In [26]:

```python
# https://www.konga.com/category/accessories-computing-5227
# https://www.konga.com/category/phones-tablets-5294
# https://www.konga.com/category/electronics-5261
# https://www.konga.com/category/konga-fashion-1259
# https://www.konga.com/category/home-kitchen-602
# https://www.konga.com/category/baby-kids-toys-8
# https://www.konga.com/category/other-categories-5971
```

Get each page section link

In [27]:

```python
div_links=page.find("div",{"class":"_0d327_3ftRW"})
```

In [28]:

```
qq=div_links.find_all('a')[1:]
```

In [29]:

```
" ".join(qq[0]['href'].split('/')[-1].split('-')[:-1])
```

Out[29]:

```
'accessories computing'
```

In [30]:

```
section_url=[f"https://www.konga.com/category/{qq[i]['href'].split('/')[-1]}" for i in rang
section_url
```

Out[30]:

```
['https://www.konga.com/category/accessories-computing-5227',
 'https://www.konga.com/category/phones-tablets-5294',
 'https://www.konga.com/category/electronics-5261',
 'https://www.konga.com/category/konga-fashion-1259',
 'https://www.konga.com/category/home-kitchen-602',
 'https://www.konga.com/category/baby-kids-toys-8',
 'https://www.konga.com/category/other-categories-5971']
```

Get the data of all section through all the pages available

In [31]:

```
section_urls=[f"https://www.konga.com/category/{qq[i]['href'].split('/')[-1]}" for i in ran

for url in section_urls:
    section_data=[]
    data=all_page_data(url)
    section_data +=[data]
    section_name=" ".join(url.split('/')[-1].split('-')[:-1])
    print("\n RUNNING SECTION : "+section_name)

#   Store each section data into a csv file with the specific section name
    pd.concat([ pd.DataFrame(j) for j in section_data ]).to_csv( f"{section_name}.csv" ,ind
```

```
 RUNNING SECTION : accessories computing

 RUNNING SECTION : phones tablets

 RUNNING SECTION : electronics

 RUNNING SECTION : konga fashion

 RUNNING SECTION : home kitchen

 RUNNING SECTION : baby kids toys

 RUNNING SECTION : other categories
```

In [37]:

```python
# View one page section data
pd.read_csv("electronics.csv").head(20)
```

Out[37]:

	off %	Description	Price	store name	number of review	available at
0	10%	iTec 32" Hd Led With Free Wall Bracke...	₦37,690	ashour-fzco	0	Lagos,Abuja
1	6%	iTec 32" HD LED with Free Wall Bracke...	₦39,500	ashour-fzco	0	Lagos
2	10%	Video Camera Camcorder Tripod with Ph...	₦4,500	it	8 Reviews	NaN
3	50%	3.5mm Mini Studio Hands Free Micropho...	₦2,000	it	21 Reviews	NaN
4	22%	Nsonic 24-inch Full HD LED TV + Free ...	₦35,000	world	0	NaN
5	33%	Samsung 32" High Definition Flat TV	₦60,000	uchegold76	0	NaN
6	60%	Imagine 3 Pole 3.5mm Mini Studio Hand...	₦2,000	it	0	NaN
7	44%	Lightweight Smartphone And Video Came...	₦4,500	resources	5 Reviews	NaN
8	13%	7ft 2m Photo Studio Video Camera Ligh...	₦7,000	it	1 Review	NaN
9	50%	Imagine 3 Pole 3.5mm Mini Studio Hand...	₦1,500	it	0	NaN
10	17%	iTec 42" Smart Hd With Free Wall Brac...	₦83,275	ashour-fzco	0	Lagos
11	23%	Polystar Portable Radio Dvd System Pv...	₦15,420	michi	0	Lagos
12	NaN	Amani 32" Full HD LED TV	₦38,500	electronic	0	NaN
13	25%	iTec 40" FHD LED Television	₦60,000	store	0	NaN
14	50%	Imagine 3.5mm Audio Stereo Splitter 4...	₦1,500	it	0	NaN
15	9%	iTec 50" Uhd Smart Led Television + F...	₦124,155	ashour-fzco	0	Lagos
16	14%	Bluetooth Earphones - Black	₦2,999	sedge	0	NaN
17	13%	Djack Powerful Bluetooth Home Theatre...	₦16,500	stores	0	PH
18	NaN	Hisense 49-Inch Smart Full HD LED TV ...	₦145,000	store	0	NaN
19	27%	iTec 43' HD LED Tv + Free Wall Bracke...	₦79,730	konga	0	Lagos

In [32]:

```python
driver.quit()
```