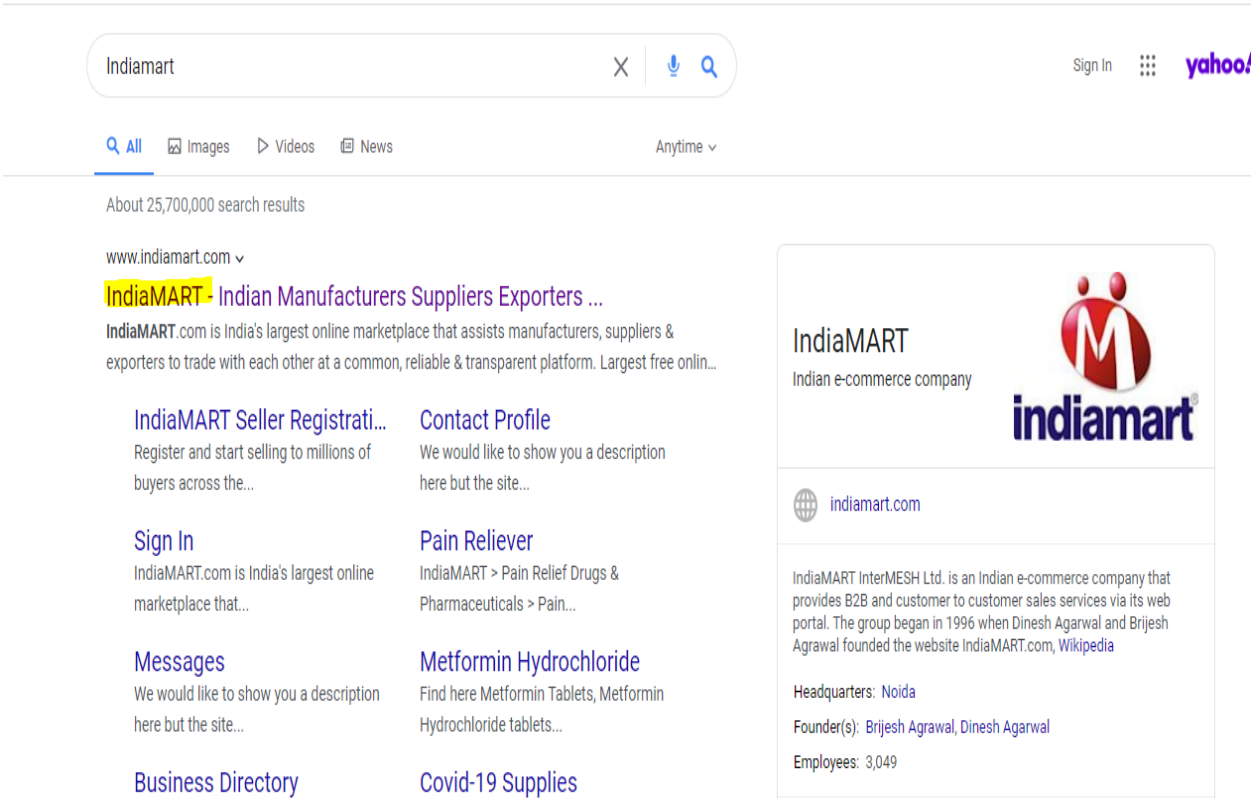
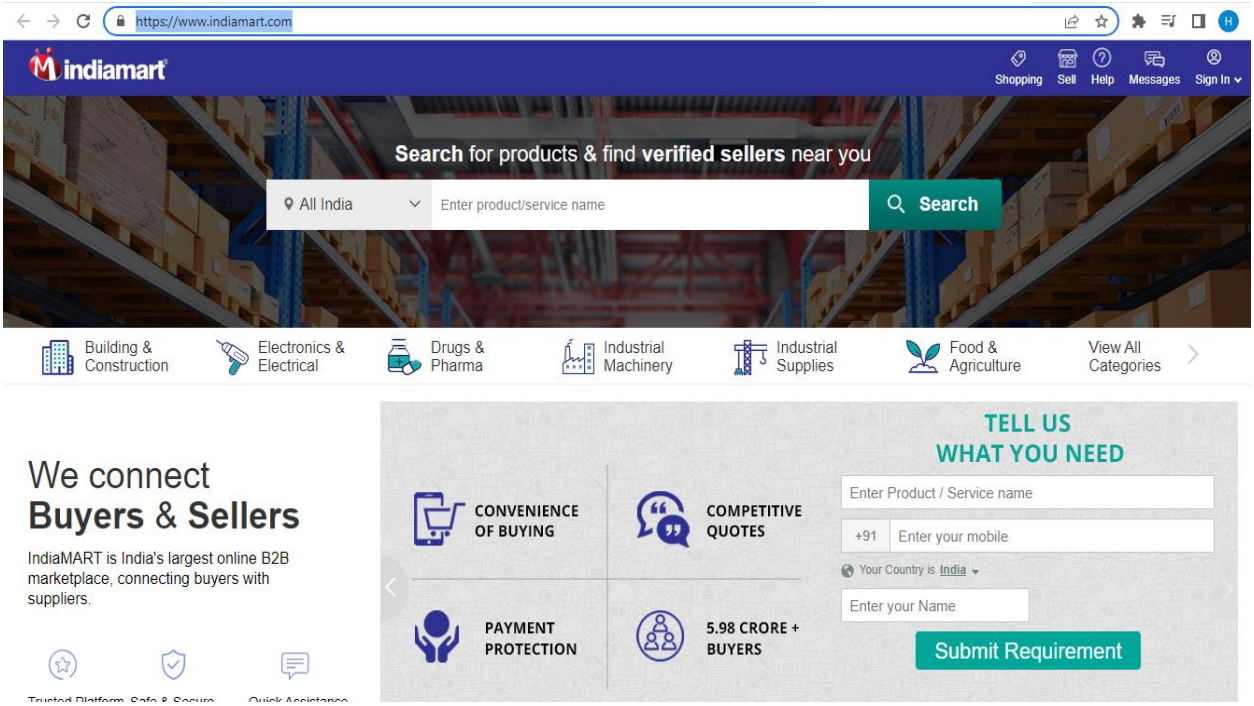
TITLE OF THE PROJECT: EXTRACT DATA FROM INDIAMART WEBSITE USING PYTHON

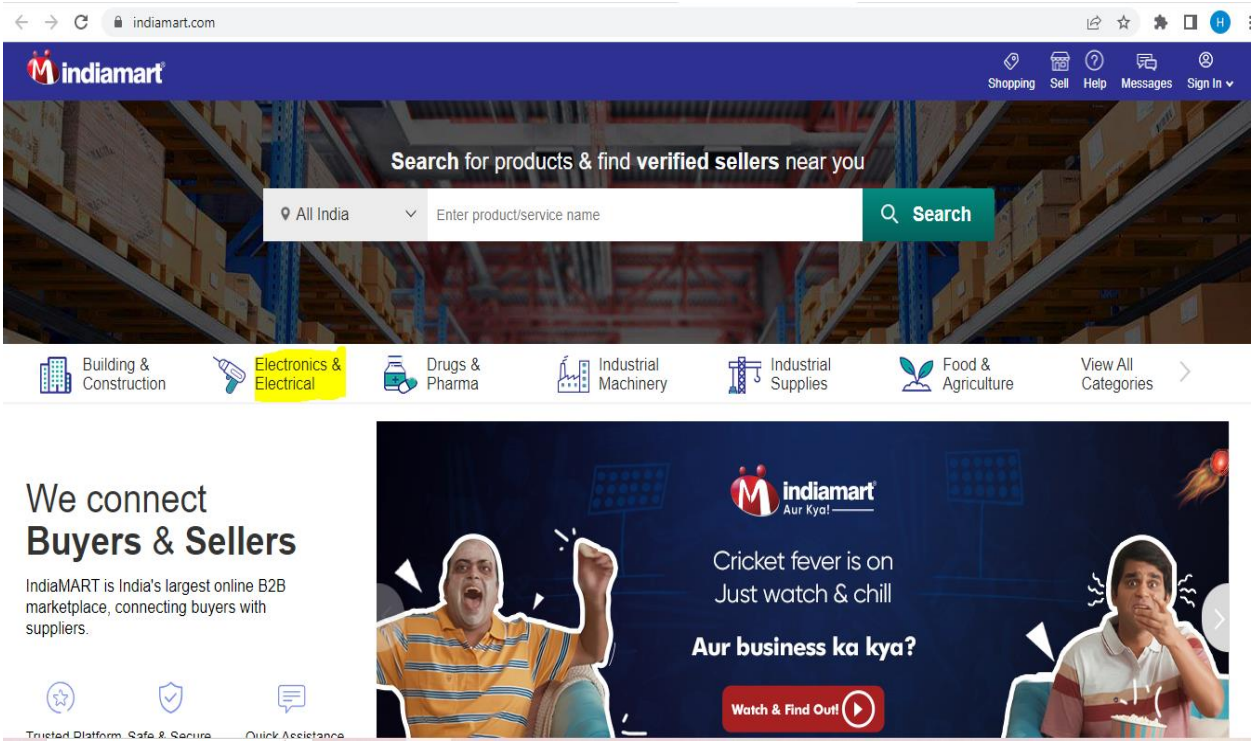
- Go to Indiamart website



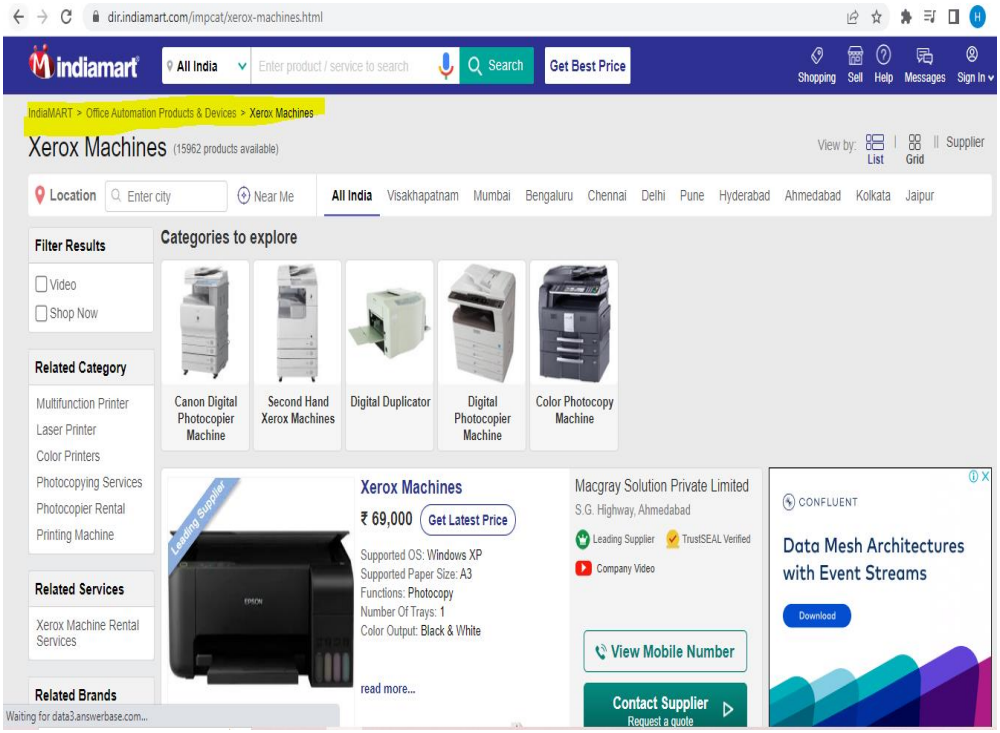
- Click on the first link:
- https://www.indiamart.com/



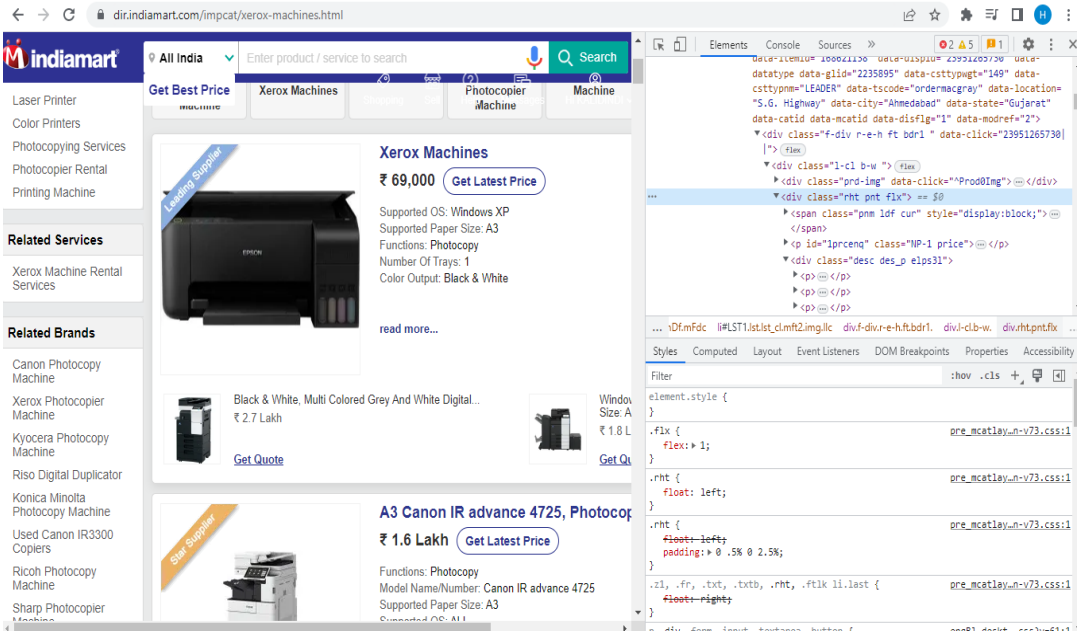
- Click on electronics and electrical:



- Click on xerox machines:



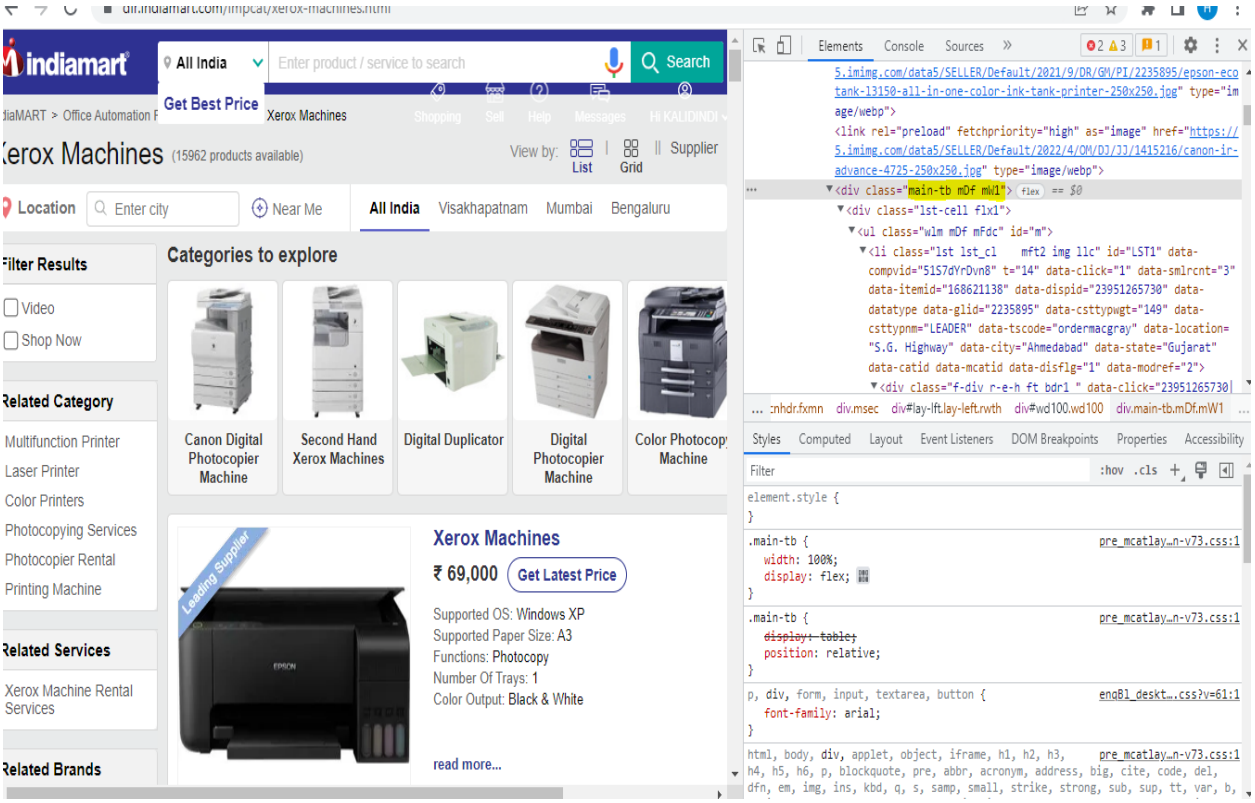
- Click on Inspect:



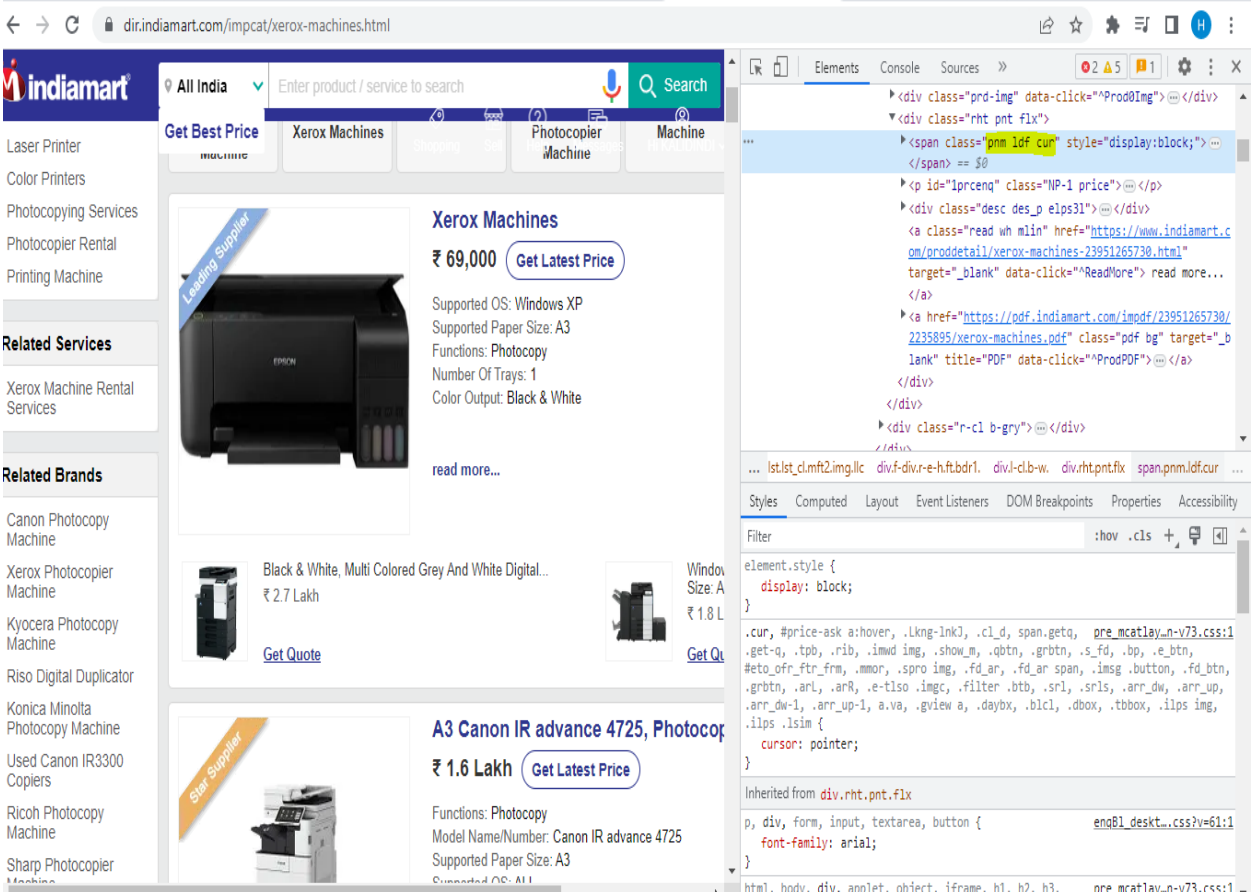
- Now we have to extract the data.
- First, we have to import the below libraries:

1. import pandas as pd
2. import requests
3. from bs4 import BeautifulSoup

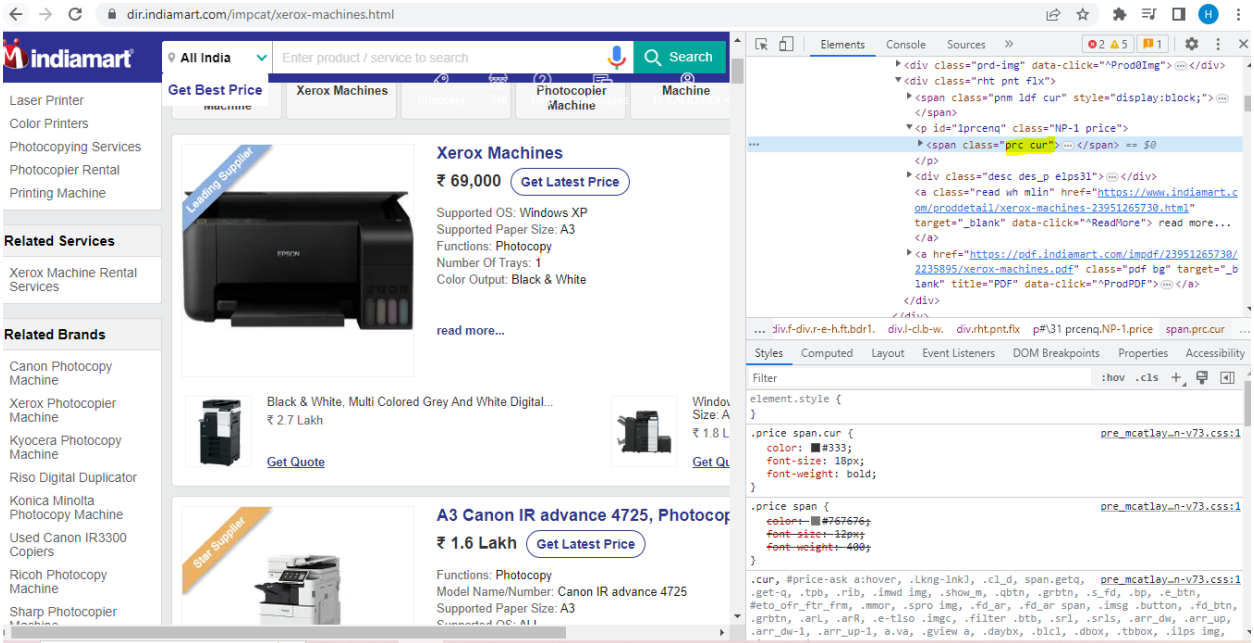
- Also collect the required information to be extracted from the page:
- First: select the box class:



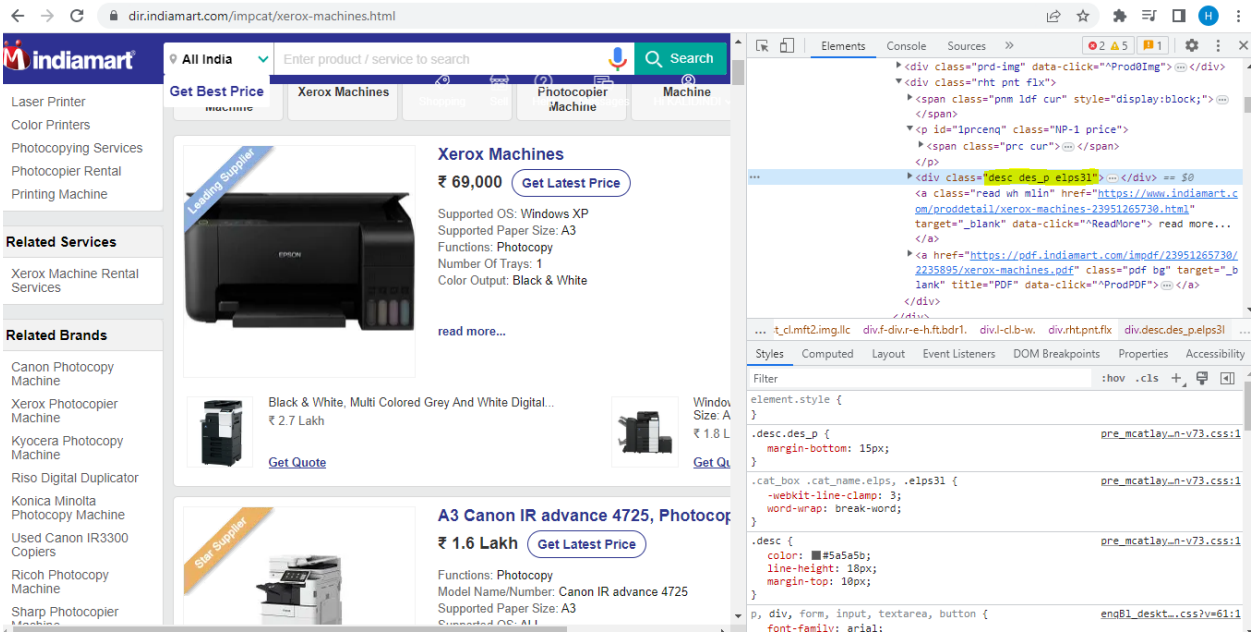
- Name of the product-



- Price of the product-



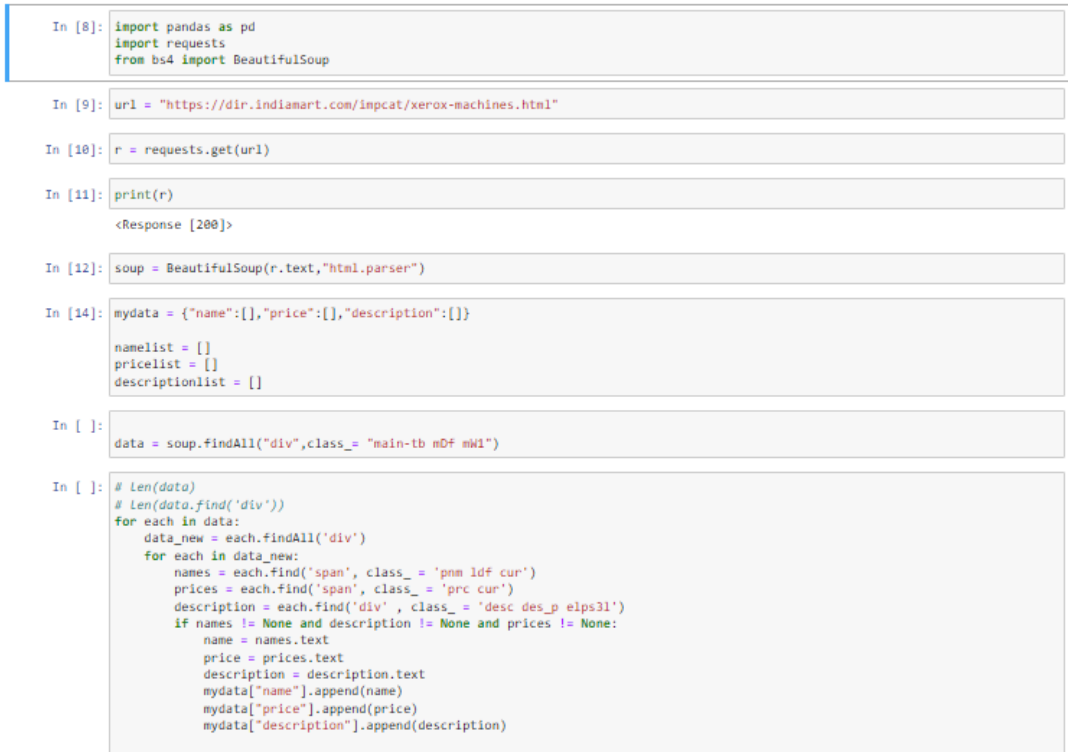
- Finally: The description of the product-



- Once we have the class information use the below code to extract the information:

```
url = https://dir.indiamart.com/impcat/xerox-machines.html

r = requests.get(url)

print(r)

soup = BeautifulSoup(r.text,"html.parser")

mydata = {"name":[],"price":[],"description":[]}

namelist = []

pricelist = []

descriptionlist = []

data = soup.findAll("div",class_= "main-tb mDf mW1")

# len(data)
```

```
# len(data.find('div'))
for each in data:
    data_new = each.findAll('div')
    for each in data_new:
        names = each.find('span', class_ = 'pnm ldf cur')
        prices = each.find('span', class_ = 'prc cur')
        description = each.find('div' , class_ = 'desc des_p elps3l')
        if names != None and description != None and prices != None:
            name = names.text
            price = prices.text
            description = description.text
            mydata["name"].append(name)
            mydata["price"].append(price)
            mydata["description"].append(description)
```



```
In [8]:
import pandas as pd
import requests
from bs4 import BeautifulSoup

In [9]:
url = "https://dir.indiamart.com/impcat/xerox-machines.html"

In [10]:
r = requests.get(url)

In [11]:
print(r)

<Response [200]>

In [12]:
soup = BeautifulSoup(r.text,"html.parser")

In [14]:
mydata = {"name":[],"price":[],"description":[]}

namelist = []
pricelist = []
descriptionlist = []

In [ ]:

data = soup.findAll("div",class_= "main-tb mDf mW1")

In [ ]:
# len(data)
# len(data.find('div'))
for each in data:
    data_new = each.findAll('div')
    for each in data_new:
        names = each.find('span', class_ = 'pnm ldf cur')
        prices = each.find('span', class_ = 'prc cur')
        description = each.find('div' , class_ = 'desc des_p elps3l')
        if names != None and description != None and prices != None:
            name = names.text
            price = prices.text
            description = description.text
            mydata["name"].append(name)
            mydata["price"].append(price)
            mydata["description"].append(description)
```

- After extracting the data remove the duplicates and use the below code to load the data in csv file and set the location for the file storage:

print(mydata)

df = pd.DataFrame.from_dict(mydata)

data = df.drop_duplicates()

data.head()

data.to_csv('C:/Users/DELL/Desktop/Harika_DE_projects/beautifulSoupdataextractproject.csv')

```
In [ ]: # print(mydata)

        df = pd.DataFrame.from_dict(mydata)

In [ ]: data = df.drop_duplicates()

In [ ]: # data.head()
        data.to_csv('C:/Users/DELL/Desktop/Harika_DE_projects/beautifulSoupdataextractproject.csv'
```

- Open and check the csv file: