Excel and Google Sheets for Web Scraping

Dr. Anil Shrestha
Undersecretary (Account)
Financial Administration Section
National Statistics Office

Pros and Cons

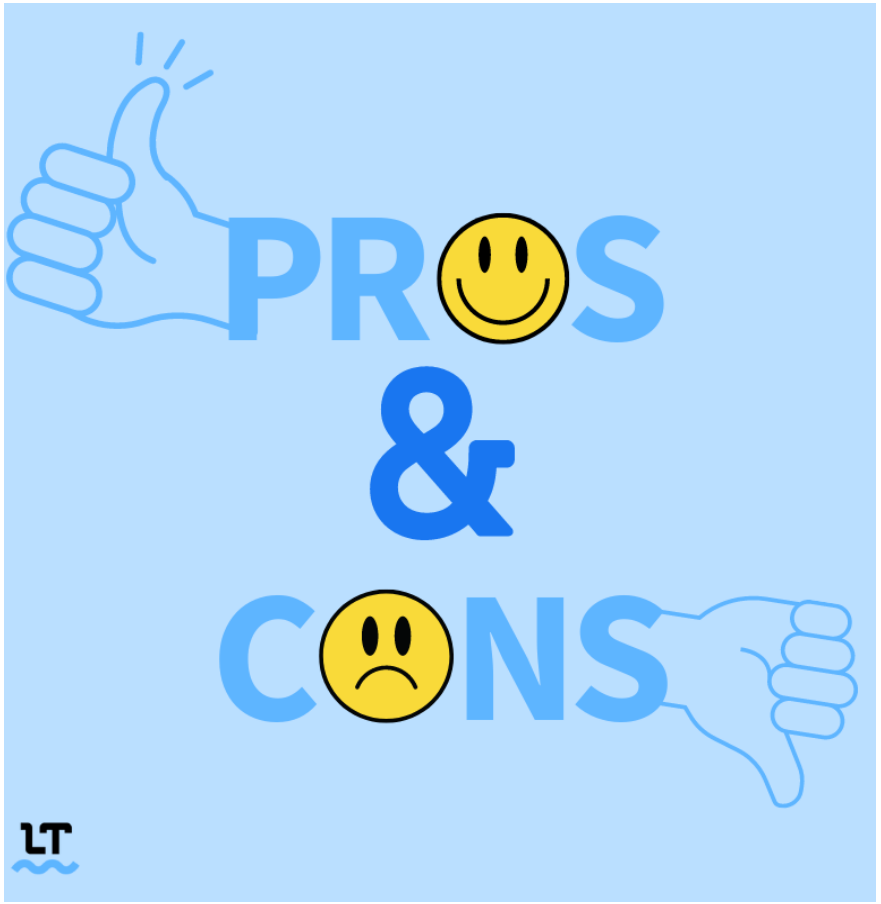
Pros & Cons

Pros

- User-friendly and accessible, especially for non-coders.
- No need to write complex code, saving time and resources.
- Built-in features make cleaning and organizing data simple and quick.

Cons

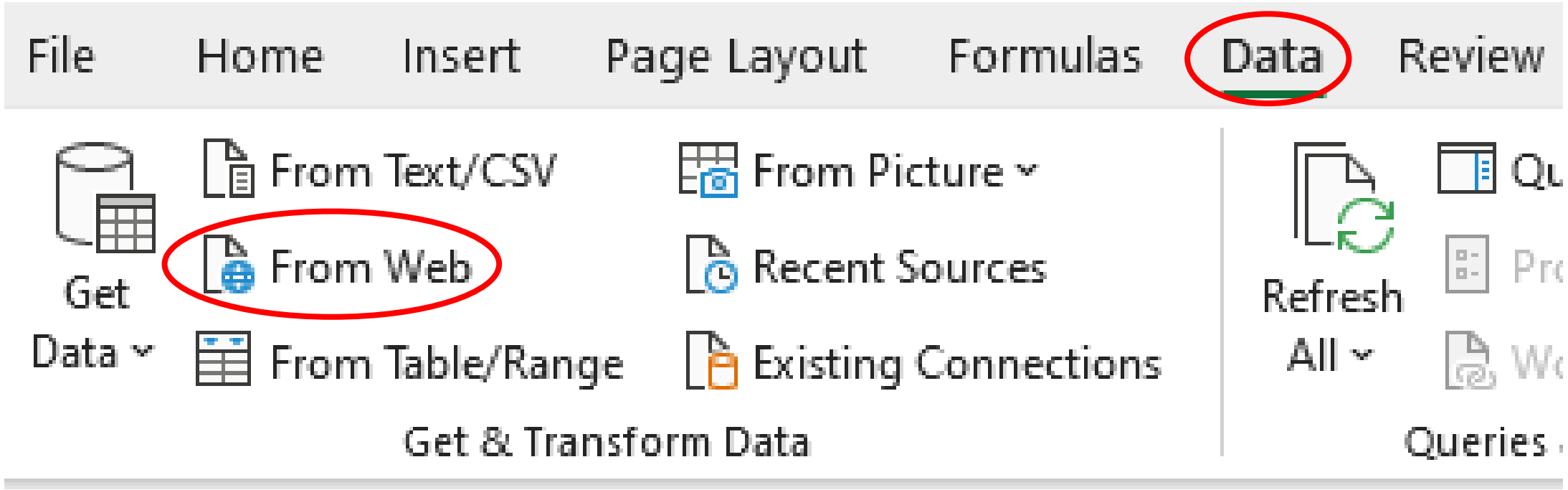
- May not be suitable for large-scale or complex data extraction projects.
- Limited customization and control compared to coding from scratch.
- Not suitable for dynamic websites.
- Could be slow for processing large amounts of data.

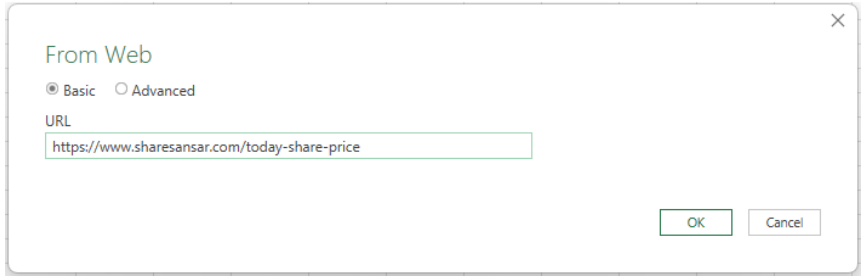


Excel’s Web Query

Web Query in Excel

- Let's scrap share price from https://www.sharesansar.com/today-share-price using Excel's web query feature.





Navigator

Select multiple items

Display Options

- HTML Tables [1]
  - Table 1
- Suggested Tables [10]
  - Table 2
  - Table 3
  - Table 4
  - Table 5
  - Table 6
  - Table 7
  - Table 8
  - Table 9
  - Table 10
  - Table 11
- Text [2]
  - HTML Code
  - Displayed Text

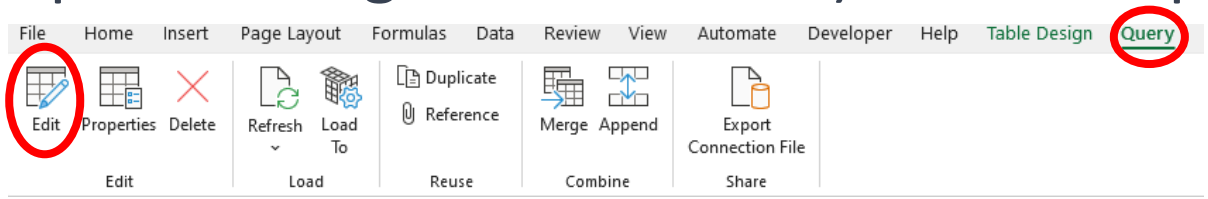
Table View | Web View

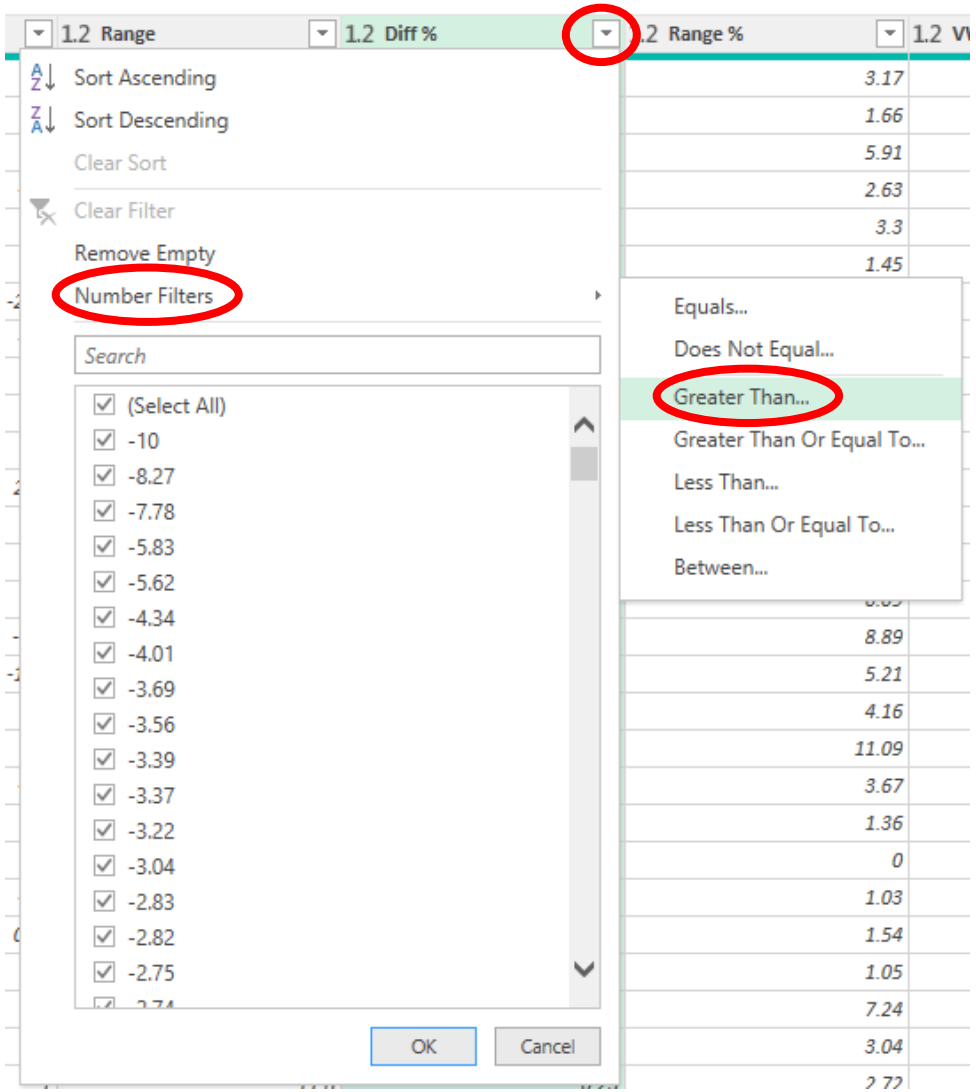
Table 1

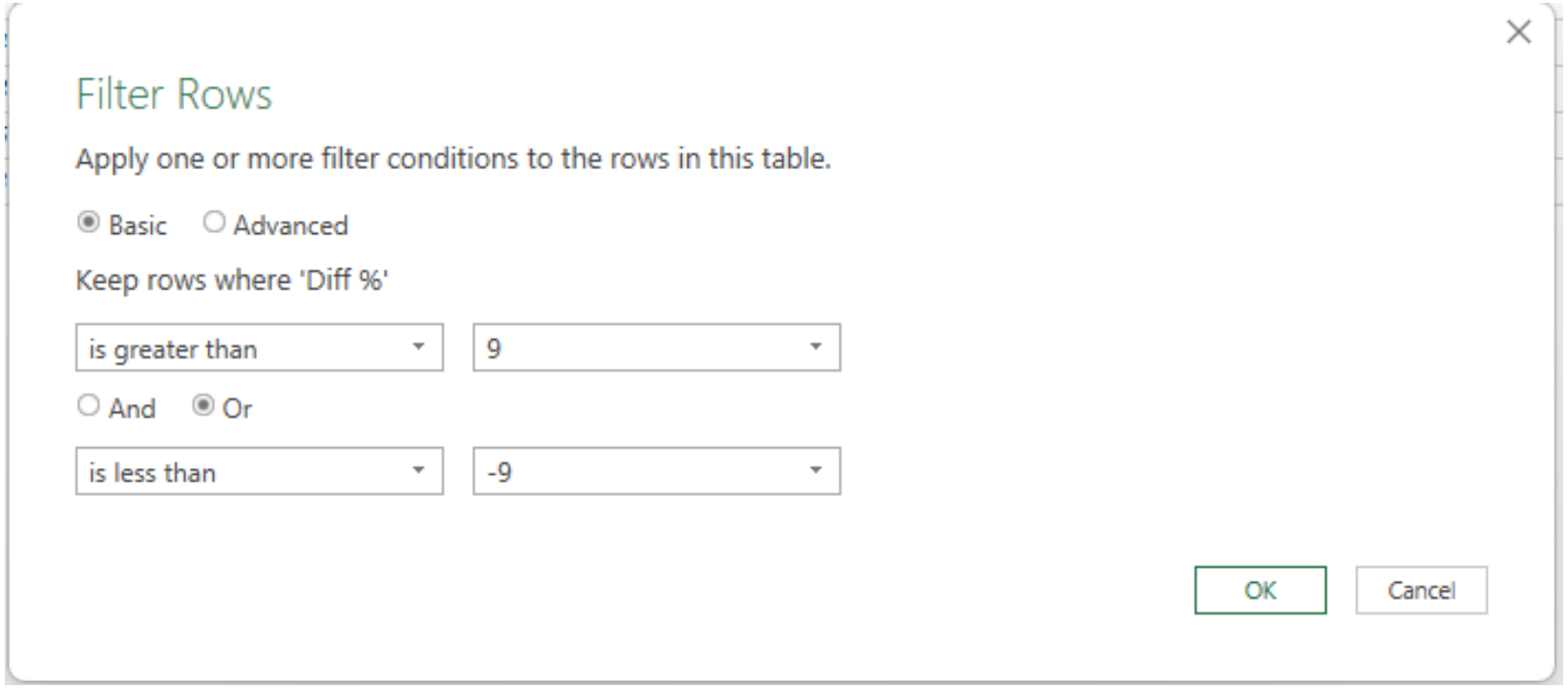
S.No	Symbol	Conf.	Open	High	Low	Close
1	ACLBSL	50.35	1043	1043	1011	
2	ADBL	56.45	269	269	264.6	
3	AHL	53.25	516.1	519.9	490.9	
4	AHPC	59.44	159	159.9	155.8	1
5	AKJCL	55	206.2	213	206.2	
6	AKPL	59.46	168	168	165.6	
7	ALBSL	47.03	1032	1039.9	1009	10
8	ALICL	57.79	568	588	567.1	5
9	ANLB	42.77	2345	2370	2301.3	
10	API	60.37	176	176	172.5	
11	AVYAN	57.1	988	989.4	945.1	
12	BARUN	66.38	288.7	311.1	283.7	
13	BBC	63.94	3985	4145.2	3975	
14	BEDC	46.89	406	426	406	
15	BFC	58.08	455.9	470	443	
16	BGWT	42.36	1208.4	1208.4	1109.7	
17	BHDC	44.02	580	599.8	570.1	5
18	BHL	63.44	361.1	376	361	3
19	BHPL	58.98	640	711	640	
20	BNHC	45.2	597.3	619	597.1	6
21	BNT	57.54	12901	13000	12826	1

Add Table Using Examples | Load | Transform Data | Cancel

Suppose we are only interested in upper and lower circuit stocks (i.e. price change >9% or <-9%). Let's scrap the data based on these rules.






Filter Rows
Apply one or more filter conditions to the rows in this table.
Basic
Advanced
Keep rows where 'Diff %'
is greater than
9
And
Or
is less than
-9
OK
Cancel

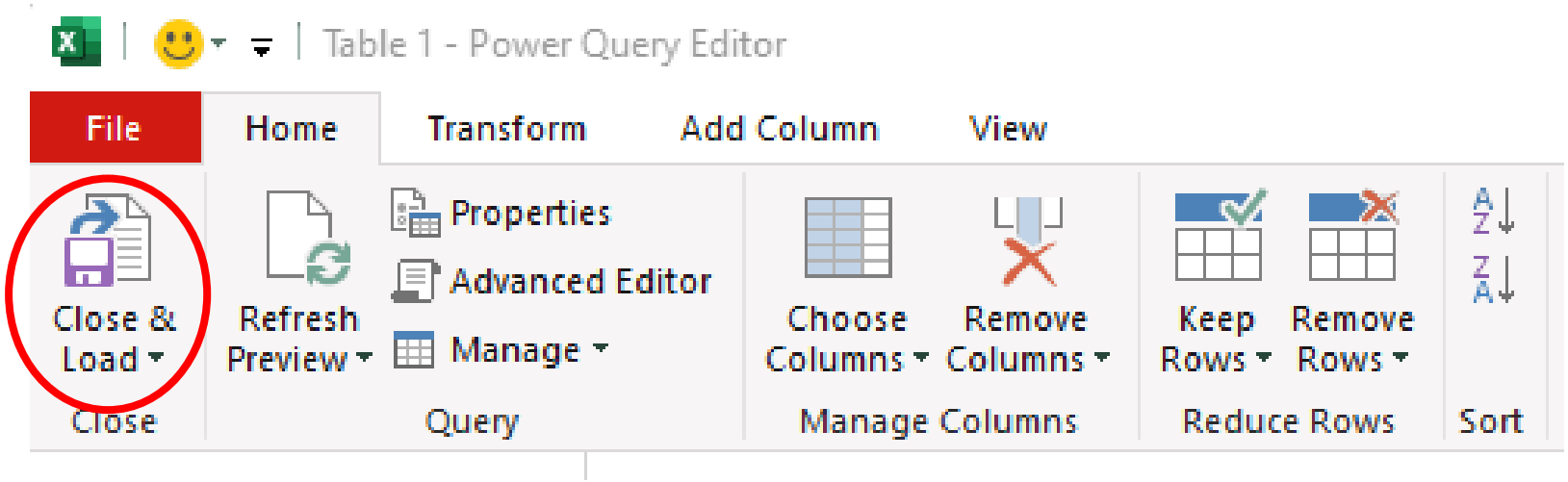


Table 1 - Power Query Editor
File
Home
Transform
Add Column
View
Close & Load
Refresh Preview
Properties
Advanced Editor
Manage
Choose Columns
Remove Columns
Keep Rows
Remove Rows
Close
Query
Manage Columns
Reduce Rows
Sort

Web scraping Google Sheets

Benefits of using Google Sheets

- Faster and simpler than excel web scraping.
- Multiple built-in web scraping functions (e.g., IMPORTDATA, IMPORTHTML, IMPORTXML)
- Stores in the cloud and easier to share/collaborate.
- Auto updating and no need to manually refresh like in Excel.

Let’s scrap share price from https://www.sharesansar.com/today-share-price using Google Sheets.

- Open any browser and enter *sheet.new* in the address bar.
- Log in with your google account (if not logged in).
- Give a name to the sheet. (Sheet will be stored in google drive)

Let's filter the upper and lower circuit stocks (i.e. price change >9% or <-9%). For this, we need to embed IMPORTHTML function inside QUERY function.

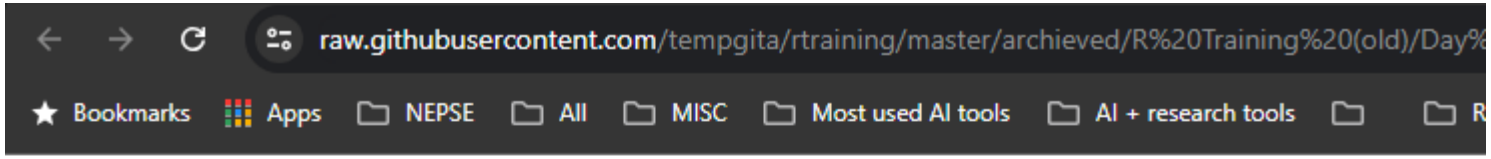
```
=query(importhtml("https://www.sharesansar.com/today-share-price","table",1),"select * where Col15 > 9 or Col15 <-9")
```

B	C	D	E	F	G	H	I	J	K	L	M	N	O	
Symbol	Conf.	Open	High	Low	Close	VWAP	Vol	Prev. Close	Turnover	Trans.	Diff	Range	Diff %	Ra
BHPL	58.98	640.00	711.00	640.00	710.00	677.50	46,608.00	647.00	31,577,062.70	528	63	71	9.74	

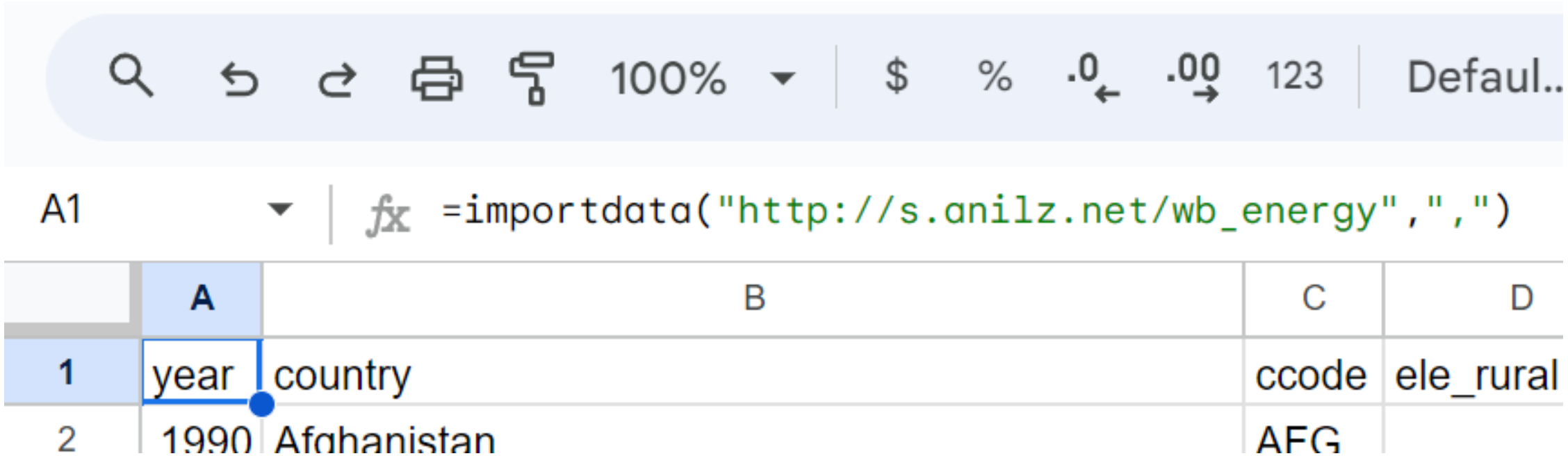
Scraping CSV data

Sometimes, data is available in CSV format. We can import CSV data directly to Google Sheet using **IMPORTDATA** function.

For example, open the URL http://s.anilz.net/wb_energy




```
year,country,ccode,ele_rural,ele_total,ele_urban,en_int,ren_ele,ren_con,tot_ele,tfec
1990,Afghanistan,AFG,,0.01,52.03697586,1.884112773,764,6312.392,1128,39639.42002
1990,Albania,ALB,100,100,100,7.912243196,2848,20429.18,3296,80057.64499
1990,Algeria,DZA,96.39231475,98.27137756,100,3.500934776,135,811.7773,16104,458040.4417
1990,American Samoa,ASM,,,,,0,0,100,306
1990,Andorra,AND,100,100,100,,120,952.145,120,6670.695
1990,Angola,AGO,7.518615066,11.39780807,22.68237495,4.605299718,725,135443.7,841,187451.7027
1990,Anguilla,AIA,,89.1986618,89.1986618,,0,1.827,16.7,615.397
1990,Antigua and Barbuda,ATG,76.96147746,85.12319946,100,3.953882216,0,0,95,2551.9
1990,Argentina,ARG,90.64082336,90.64082336,90.64082336,5.439096615,17983,105714.2,50740,1184750.603
1990,Armenia,ARM,95.78970448,97.68037415,98.59397888,24.37219694,1555,5502.305,10362,259671.993
1990,Aruba,ABW,76.75690793,88.44535065,99.9854126,,0,8.26657,338,3075.96657
1990,Australia,AUS,100,100,100,7.416865313,14898,176729.7,154287,2206479.774
1990,Austria,AUT,100,100,100,4.360147231,32635,191963.7,49296,763707.7403
1990,Azerbaijan,AZE,92.47360267,95.87023163,98.79302979,15.56773215,1658,4783.298,23152,662449.1031
1990,"Bahamas, The",BHS,82.02916466,90.79119873,93.00390625,4.3429759,0,0,950,15588.9
1990,Bahrain,BHR,,,,12.48408539,0,0,7989,43713.16463
1990,Bangladesh,BGD,,8.544374466,65.98397827,3.899469758,884,285257.2,7732,398042.7185
1990,Barbados,BRB,,,,4.645774512,0,1805.187,468,9529.98721
1990,Belarus,BLR,100,100,100,23.12784454,20,8598.692,39526,1052010.83
```

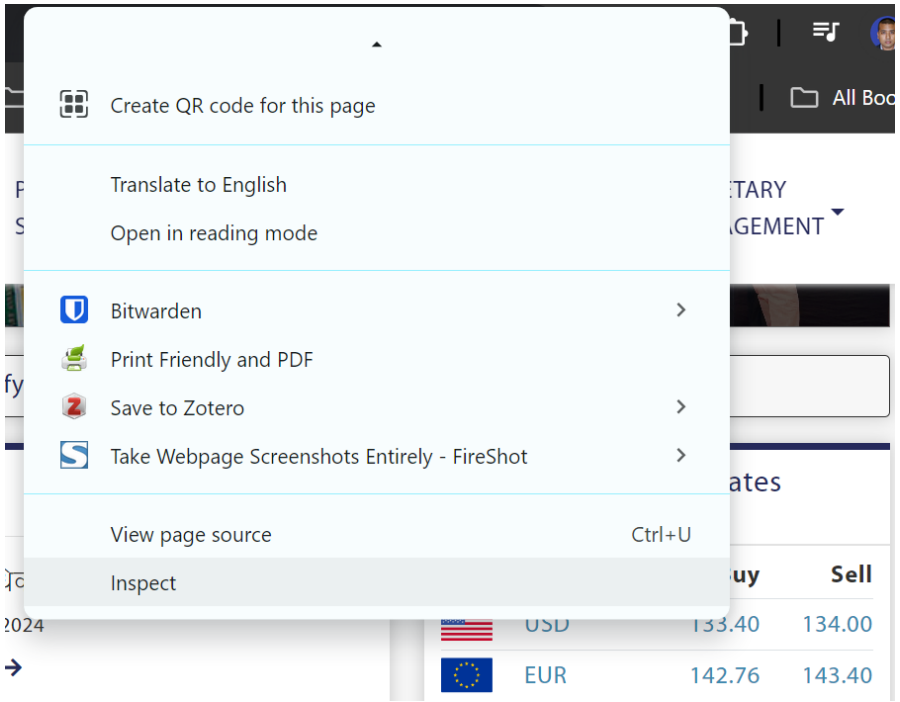

100%
$
%
.0
.00
123
Defaul..
A1
=importdata("http://s.anilz.net/wb_energy",",")
A
B
C
D
1
year
country
ccode
ele_rural
2
1990
Afghanistan
AFG

Advance web scraping Google Sheets

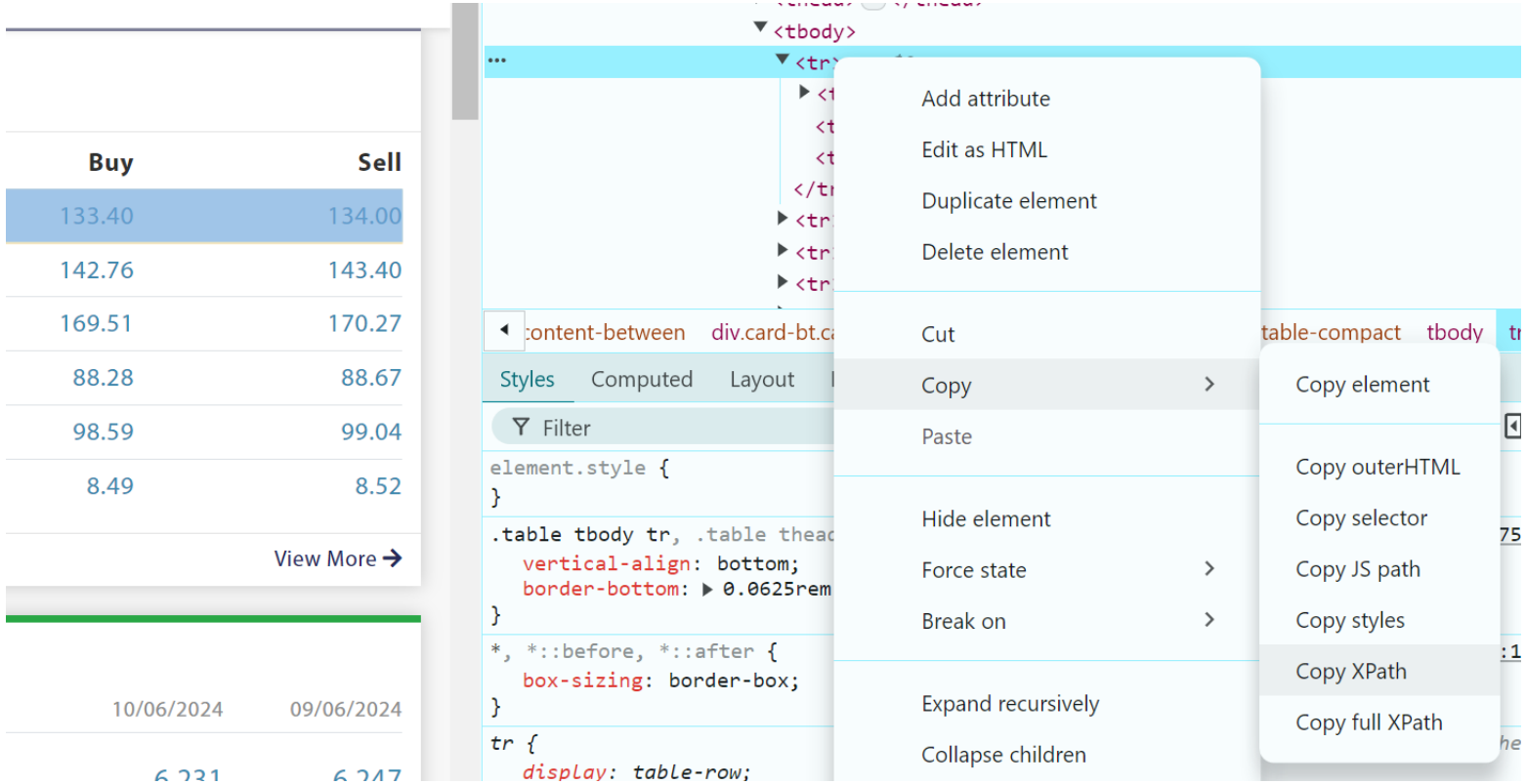
Scraping specific values

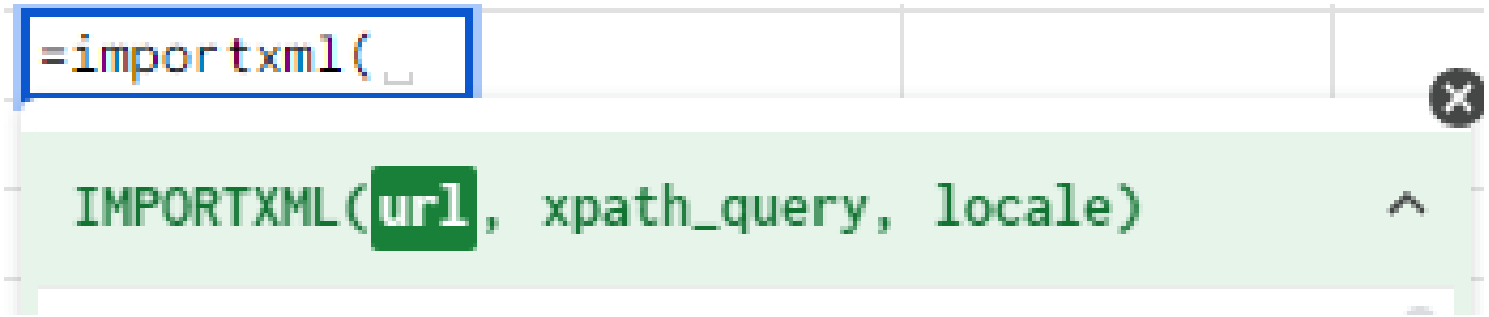
If we are interested in a specific values than a set of data, we can use **IMPORTXML** function.

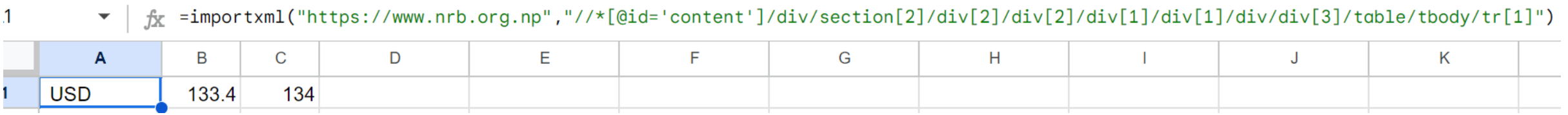
Suppose we are interested in Buy and Sell rates of USD from https://www.nrb.org.np/. To extract the Buy and Sell rates of USD, we need to extract the **XPATH** of the required values to be used in **IMPORTXML** function.



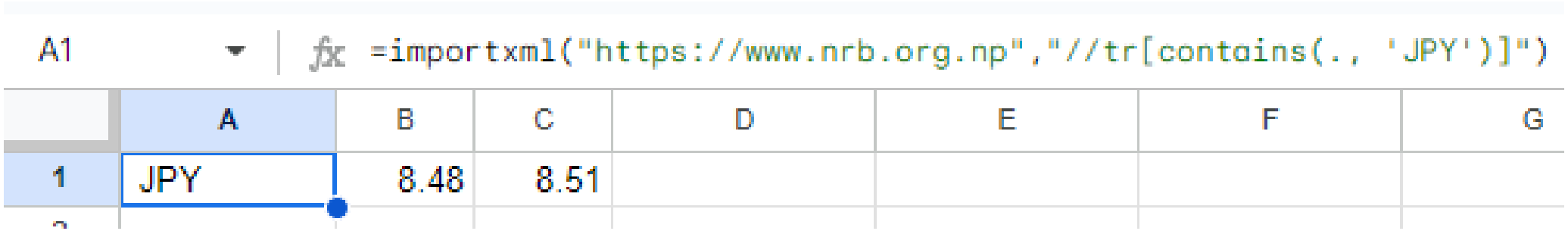
<= Can also use **Ctrl + Shift + C** to inspect element.







Instead of a fixed **XPATH**, we can also use a conditional **XPATH** for dynamically scraping exchange rate for a particular currency.



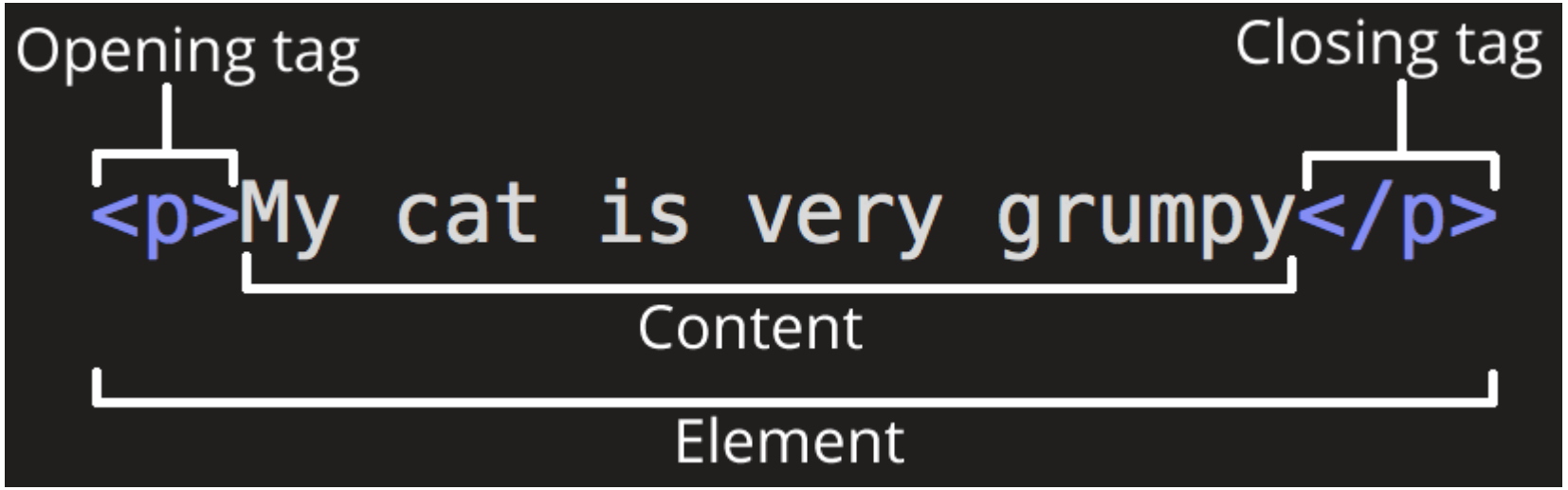
Understanding XML & XPATH

EXtensible Markup Language

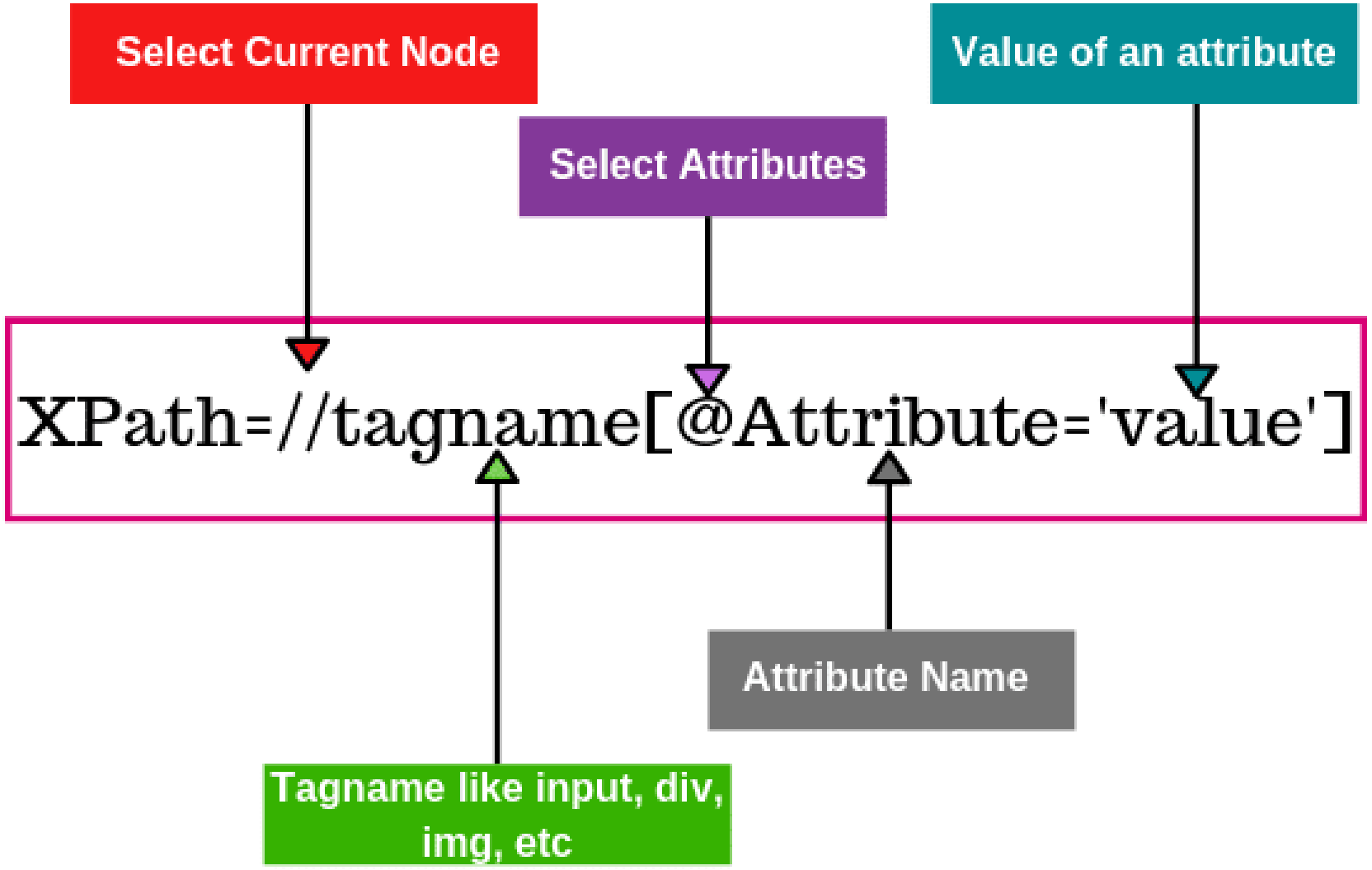
- Uses tags (e.g., <tag>data</tag>) to define and organize data elements.
- XML organizes data in tree-like structure.
- It is extensible because we can use any tag.
- It is both human and machine readable.

```
<bookstore>
  <book category="cooking">
    <title lang="en">Everyday Italian</title>
    <author>Giada De Laurentiis</author>
    <year>2005</year>
    <price>30.00</price>
  </book>

  <book category="web">
    <title lang="en">XQuery Kick Start</title>
    <author>James McGovern</author>
    <author>Per Bothner</author>
    <year>2003</year>
    <price>49.99</price>
  </book>
</bookstore>
```


Opening tag
Closing tag
My cat is very grumpy
Content
Element


Attribute
My cat is very grumpy


Select Current Node
Value of an attribute
Select Attributes
XPath=//tagname[@Attribute='value']
Attribute Name
Tagname like input, div, img, etc

Let’s practice XPATH

Go to http://xpather.com/ofMAZUAS and try to generate XPATH for the following.

- Select <country> tags.
- Select <place> where book category is “cooking”.
- Select <city> regardless of book category.
- Select <year> and <price> tags.

```
<bookstore>
  <book category="cooking">
    <title lang="en">Everyday Italian</title>
    <author>Giada De Laurentiis</author>
    <year>2005</year>
    <price>30.00</price>
    <place>
      <country>USA</country>
      <city>New York</city>
      <postalcode>558822</postalcode>
    </place>

  </book>

  <book category="web">
    <title lang="en">XQuery Kick Start</title>
    <author>James McGovern</author>
    <author>Per Bothner</author>
    <year>2003</year>
    <price>49.99</price>
    <place>
      <country>UK</country>
      <city>London</city>
    </place>
  </book>
</bookstore>
```

Select <country> tags.

```
//country
```

XML mode | Format | Save

```
<bookstore>
  <book category="cooking">
    <title lang="en">Everyday Italian</title>
    <author>Giada De Laurentiis</author>
    <year>2005</year>
    <price>30.00</price>
    <place>
      <country>USA</country>
      <city>New York</city>
      <postalcode>558822</postalcode>
    </place>
  </book>

  <book category="web">
    <title lang="en">XQuery Kick Start</title>
    <author>James McGovern</author>
    <author>Per Bothner</author>
    <year>2003</year>
    <price>49.99</price>
    <place>
      <country>UK</country>
      <city>London</city>
    </place>
  </book>
</bookstore>
```

Select <place> where book category is “cooking”.

```
//book[@category="cooking"]/place
```

XML mode | Format | Save

```
<bookstore>
  <book category="cooking">
    <title lang="en">Everyday Italian</title>
    <author>Giada De Laurentiis</author>
    <year>2005</year>
    <price>30.00</price>
    <place>
      <country>USA</country>
      <city>New York</city>
      <postalcode>558822</postalcode>
    </place>
  </book>

  <book category="web">
    <title lang="en">XQuery Kick Start</title>
    <author>James McGovern</author>
    <author>Per Bothner</author>
    <year>2003</year>
    <price>49.99</price>
    <place>
      <country>UK</country>
      <city>London</city>
    </place>
  </book>
</bookstore>
```

Select <city> regardless of book category.

```
//city
```

XML mode | Format | Save

```
<bookstore>
  <book category="cooking">
    <title lang="en">Everyday Italian</title>
    <author>Giada De Laurentiis</author>
    <year>2005</year>
    <price>30.00</price>
    <place>
      <country>USA</country>
      <city>New York</city>
      <postalcode>558822</postalcode>
    </place>
  </book>

  <book category="web">
    <title lang="en">XQuery Kick Start</title>
    <author>James McGovern</author>
    <author>Per Bothner</author>
    <year>2003</year>
    <price>49.99</price>
    <place>
      <country>UK</country>
      <city>London</city>
    </place>
  </book>
</bookstore>
```

Select <year> and <price> tags.

```
//year|//price
```

XML mode | Format | Save

```
<bookstore>
  <book category="cooking">
    <title lang="en">Everyday Italian</title>
    <author>Giada De Laurentiis</author>
    <year>2005</year>
    <price>30.00</price>
    <place>
      <country>USA</country>
      <city>New York</city>
      <postalcode>558822</postalcode>
    </place>
  </book>

  <book category="web">
    <title lang="en">XQuery Kick Start</title>
    <author>James McGovern</author>
    <author>Per Bothner</author>
    <year>2003</year>
    <price>49.99</price>
    <place>
      <country>UK</country>
      <city>London</city>
    </place>
  </book>
</bookstore>
```

Short intro to HTML

A website is made up of **HTML** code. **HTML** is a specific type of **XML** code.

Sample

```
1
2  <html>
3    <head>
4      <title>Largest companies by market cap –
           US Stock Market</title>
5      <meta charset="UTF-8">
6    </head>
7    <body>
8      <h1>Apple : 2010 Billion</h1>
9      <h2>Saudi Aramco : 1812 Billion</h2>
10     <h3>Microsoft : 1800 Billion</h3>
11     <h4>Alphabet (Google) : 1155 Billion</h4>
12     <h5>Amazon : 869 Billion</h5>
13     <b>This data is as of 21 Dec 2022.</b>
14   </body>
15 </html>
```

Ln: 1 Col: 0 size: **399** B

Output

Apple : 2010 Billion

Saudi Aramco : 1812 Billion

Microsoft : 1800 Billion

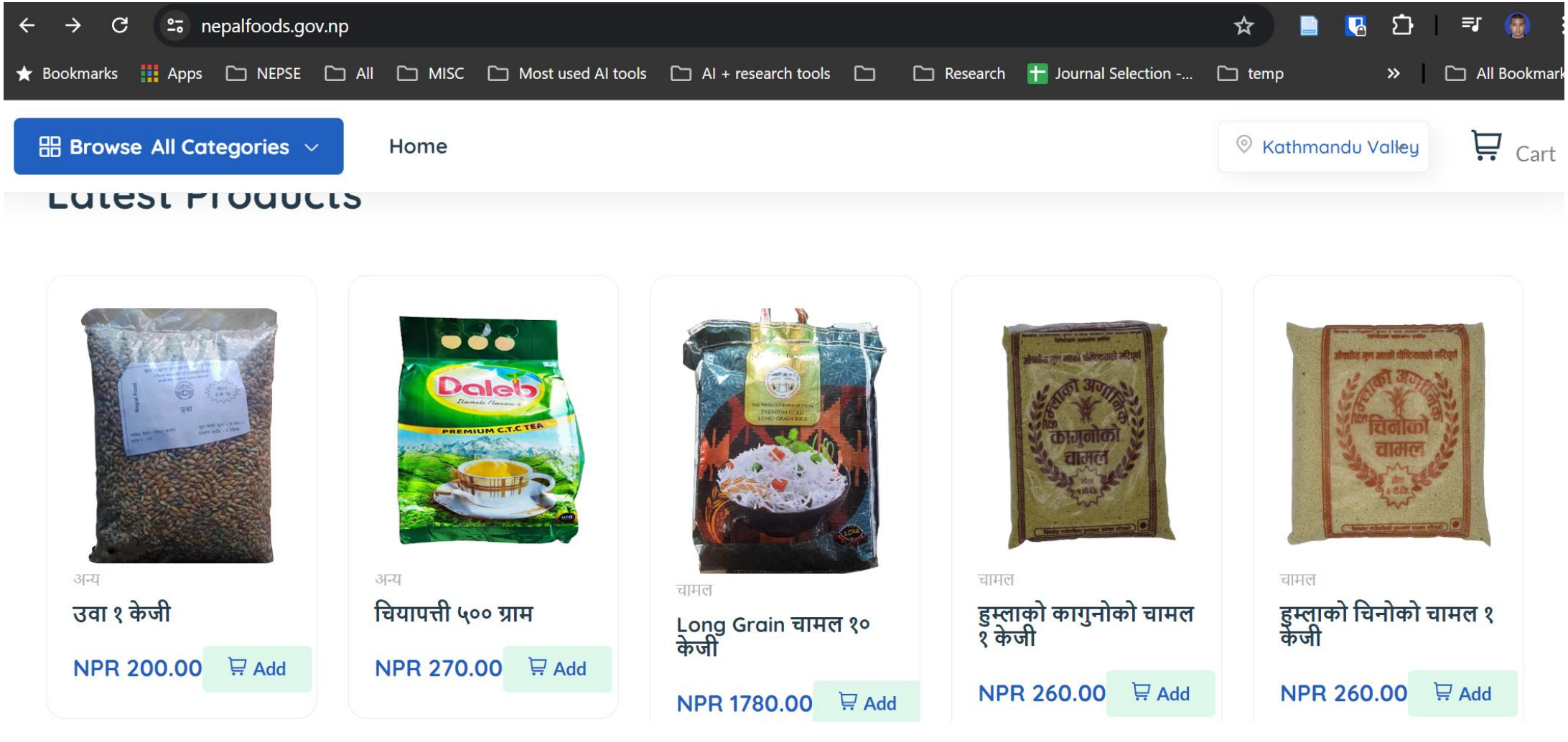
Alphabet (Google) : 1155 Billion

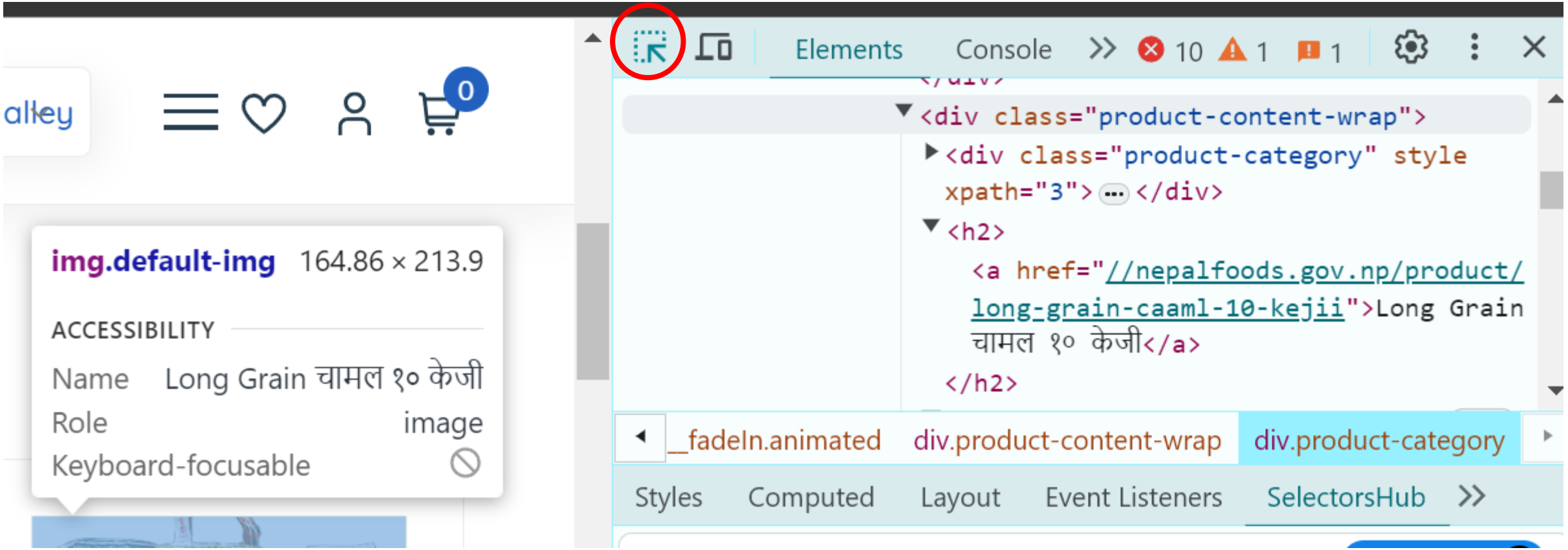
Amazon : 869 Billion

This data is as of 21 Dec 2022.

Let's scrap a shopping site

Go to https://nepalfoods.gov.np/ and scrape listed product names and prices.




alley
0
img.default-img 164.86 × 213.9
ACCESSIBILITY
Name Long Grain चामल १० केजी
Role image
Keyboard-focusable
Elements
Console
10
1
1

<div class="product-category" style
xpath="3">...
<h2>
<a href="//nepalfoods.gov.np/product/
long-grain-caaml-10-kejii">Long Grain
चामल १० केजी
</h2>
_fadeIn.animated
div.product-content-wrap
div.product-category
Styles
Computed
Layout
Event Listeners
SelectorsHub


div.product-content-wrap 244.86 × 119.77
अन्य
चियापत्ती ५०० ग्राम
NPR 270.00
Add


<div class="product-cart-wrap mb-30 wow animate__ anim
ate__fadeIn animated" data-wow-delay=".1s" style="visi
bility: visible; animation-delay: 0.1s; animation-nam
e: fadeIn;">
…

…
<h2>
b-30.wow.animate__.animate__fadeIn.animated
div.product-content-wrap
h2
a
Styles
Computed
Layout
Event Listeners
DOM Breakpoints
Properties
Filter
:hov
.cls

A1 =importxml("https://nepalfoods.gov.np/","//div[@class='product-content-wrap']")

	A	B	C	D	E
1	अन्य	उवा १ केजी	NPR 200.00 Add		
2	अन्य	चियापत्ती ५०० ग्राम	NPR 270.00 Add		
3	चामल	Long Grain चामल १० केजी	NPR 1780.00 Add		
4	चामल	हुम्लाको कागुनोको चामल १ केजी	NPR 260.00 Add		
5	चामल	हुम्लाको चिनोको चामल १ केजी	NPR 260.00 Add		
6	दाल	कर्णालीको सिमि १ केजी	NPR 240.00 Add		
7	दाल	मुसुरो दाल(सानो) १ केजी	NPR 165.00 Add		
8	चामल	अरुवा सोना मन्सुली चामल २५ केजी	NPR 1700.00 Add		
9	चामल	अरुवा मोटा चामल ३० केजी	NPR 1560.00 Add		
10	चामल	हुम्लाको कागुनोको चामल १ केजी	NPR 260.00 Add		
11	चामल	हुम्लाको चिनोको चामल १ केजी	NPR 260.00 Add		
12	अन्य	टाइमपास टाइचिन चिउरा १ केजी	NPR 100.00 Add		
13	अन्य	गहुँ आटा 5 केजी	NPR 360.00 Add		
14	तेल एवं घ्यू	भटमासको तेल १ लिटर	NPR 215.00 Add		
15	तेल एवं घ्यू	सनफ्लावर तेल १ लिटर	NPR 220.00 Add		
16	तेल एवं घ्यू	तोरीको तेल (शान्ती) १ लिटर	NPR 385.00 Add		
17	तेल एवं घ्यू	डी.डी.सी डेरी घ्यू १ लि	NPR 1160.00 Add		
18	अन्य	डी.डी.सी डेरी घ्यू १/२ लि	NPR 580.00 Add		

Tools for easier XPATH generation

XPATH finding tool

Install the following two Extensions (tools) for chrome browser for easier XPATH finding from Chrome web store (https://chromewebstore.google.com)

SelectorGadget

SelectorsHub

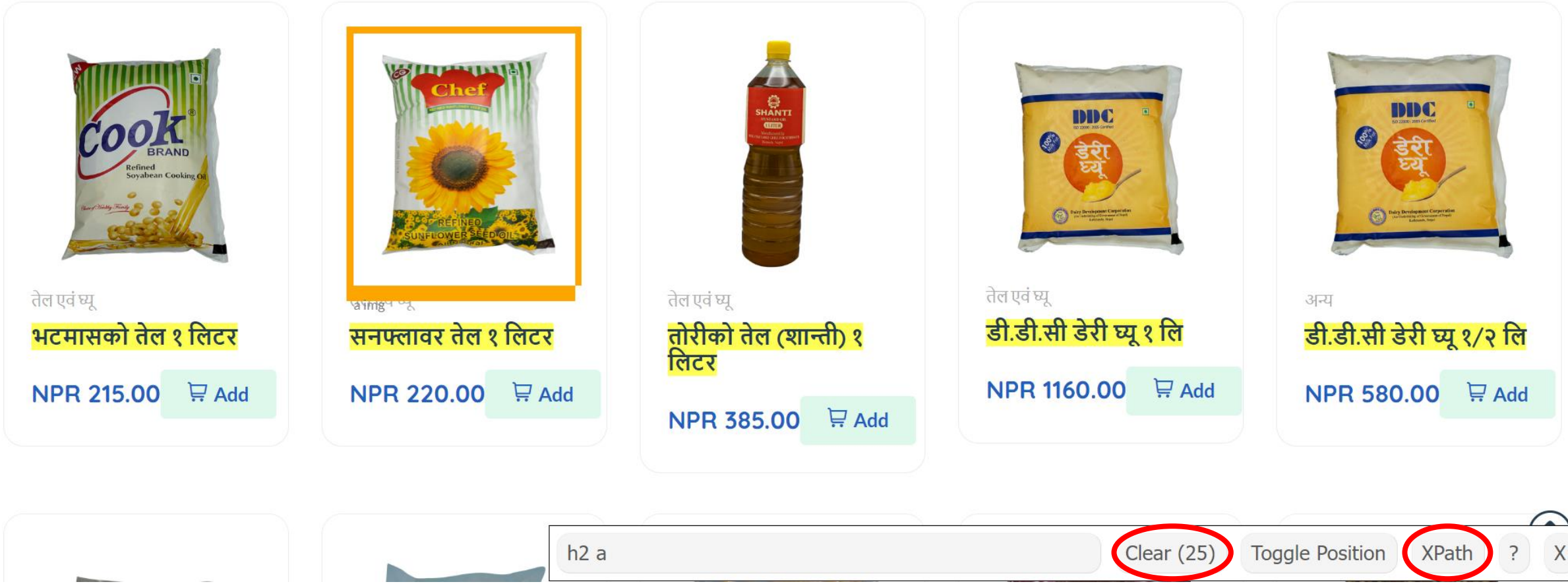
SelectorGadget

- ***SelectorGadget*** is good for visually selecting elements in group.
- Not suitable for individual element selection.

SelectorsHub

- ***SelectorsHub*** is good for visualizing individual or group of element selection.
- Flexible and can try custom XPATH with visualization feature.

Let’s use SelectorGadget to https://nepalfoods.gov.np to select elements in group.

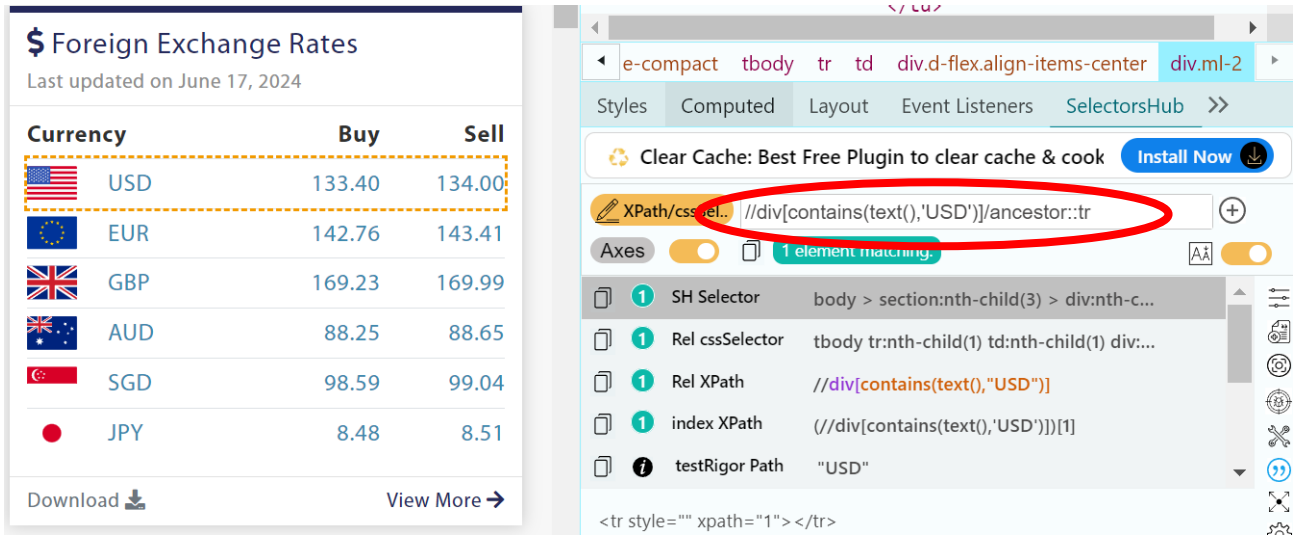


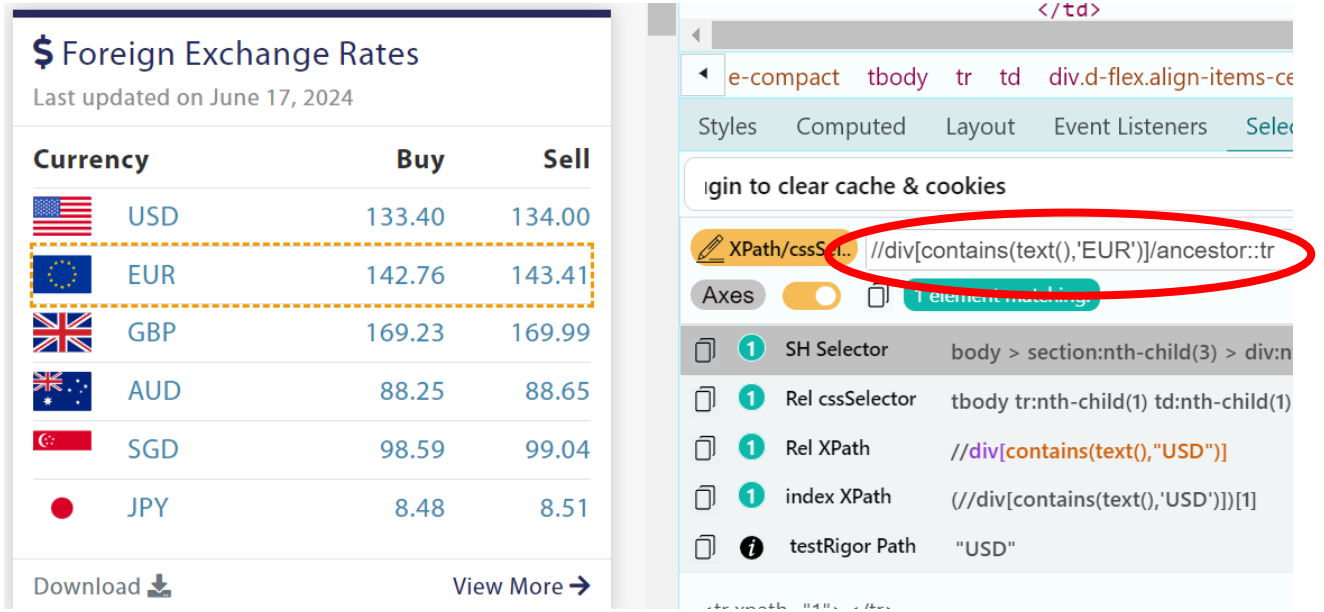
A2 | =importxml("https://nepalfoods.gov.np/",A1)

Name box (Ctrl + J)

	A	B	C
1	//*[contains(concat(" ", @class, " "), concat(" ", "product-category", " "))]	//h2//a	//*[(@id = "tab-one")]//span
2	अन्य	उवा १ केजी	NPR 200.00
3	अन्य	चियापत्ती ५०० ग्राम	NPR 270.00
4	चामल	Long Grain चामल १० केजी	NPR 1780.00
5	चामल	हुम्लाको कागुनोको चामल १ केजी	NPR 260.00
6	चामल	हुम्लाको चिनोको चामल १ केजी	NPR 260.00
7	दाल	कर्णालीको सिमि १ केजी	NPR 240.00
8	दाल	मुसुरो दाल(सानो) १ केजी	NPR 165.00
9	चामल	अरुवा सोना मन्सुली चामल २५ केजी	NPR 1700.00
10	चामल	अरुवा मोटा चामल ३० केजी	NPR 1560.00
11	चामल	हुम्लाको कागुनोको चामल १ केजी	NPR 260.00
12	चामल	हुम्लाको चिनोको चामल १ केजी	NPR 260.00
13	अन्य	टाइमपास टाइचिन चिउरा १ केजी	NPR 100.00
14	अन्य	गहुँ आटा 5 केजी	NPR 360.00
15	तेल एवं घ्यू	भटमासको तेल १ लिटर	NPR 215.00
16	तेल एवं घ्यू	सनफ्लावर तेल १ लिटर	NPR 220.00
17	तेल एवं घ्यू	तोरीको तेल (शान्ती) १ लिटर	NPR 385.00
18	तेल एवं घ्यू	डी.डी.सी डेरी घ्यू १ लि	NPR 1160.00

Let’s use SelectorsHub to https://www.nrb.org.np to make custom XPATH and visualize them.





Let’s use SelectorsHub to https://nepalfoods.gov.np to make custom XPATH and visualize them. 25 elements are matched and matched elements are highlighted in dotted lines.

A2 | =importxml("https://nepalfoods.gov.np/",A1)

	A	B	C
1	//div[@class='product-category']	//h2	//div[@class='product-price']
2	अन्य	उवा १ केजी	NPR 200.00
3	अन्य	चियापत्ती ५०० ग्राम	NPR 270.00
4	चामल	Long Grain चामल १० केजी	NPR 1780.00
5	चामल	हुम्लाको कागुनोको चामल १ केजी	NPR 260.00
6	चामल	हुम्लाको चिनोको चामल १ केजी	NPR 260.00
7	दाल	कर्णालीको सिमि १ केजी	NPR 240.00
8	दाल	मुसुरो दाल(सानो) १ केजी	NPR 165.00
9	चामल	अरुवा सोना मन्सुली चामल २५ केजी	NPR 1700.00
10	चामल	अरुवा मोटा चामल ३० केजी	NPR 1560.00
11	चामल	हुम्लाको कागुनोको चामल १ केजी	NPR 260.00
12	चामल	हुम्लाको चिनोको चामल १ केजी	NPR 260.00
13	अन्य	टाइमपास टाइचिन चिउरा १ केजी	NPR 100.00
14	अन्य	गहुँ आटा 5 केजी	NPR 360.00
15	तेल एवं घ्यू	भटमासको तेल १ लिटर	NPR 215.00
16	तेल एवं घ्यू	सनफ्लावर तेल १ लिटर	NPR 220.00
17	तेल एवं घ्यू	तोरीको तेल (शान्ती) १ लिटर	NPR 385.00
18	तेल एवं घ्यू	डी.डी.सी डेरी घ्यू १ लि	NPR 1160.00
19	अन्य	डी.डी.सी डेरी घ्यू १/२ लि	NPR 580.00
20	चामल	मार्सी चामल १ केजी	NPR 230.00

Thank you