Scraping Data from the Web

In this lab, you will learn how to scrape data from websites using browser extensions and Google Spreadsheets.

Getting Started

Despite business complexity, corporate data is surprisingly open in Russia. Photo: Shutterstock

In Russia, shell companies abound and political involvement in business creates significant risks for international businesses. Today, there is the added pressure of increasing US and EU sanctions. In this example, the International Consortium of Investigative Journalists scraped several Russian and international business registries to track ownership and money flows behind shell companies. The data was available online but there was no way to download the data from the site. Take a look at: http://www.icij.org/blog/2014/08/how-investigate-russias-shady-business-world-online

Scraping Data Using Browser Extensions

Browser Extensions or Browser Add-Ons are applications that run in your internet browser that enable you to:

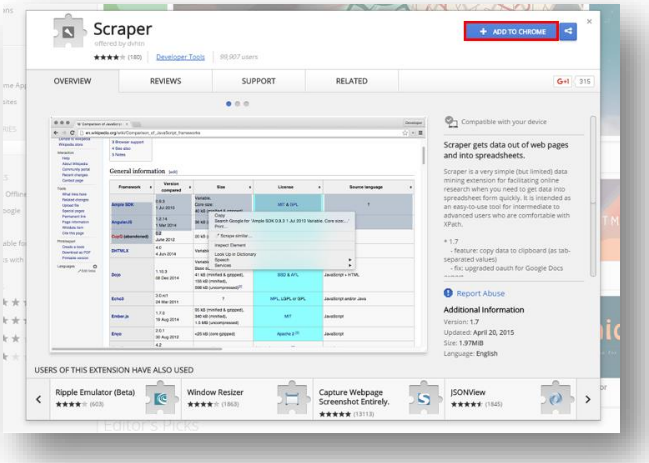
- Select data tables on a website, including specific rows and columns
- Copy data from these tables and use it in a spreadsheet application like Excel or Google Sheets

Here is an example, let's use a browser extension to scrape data about Pakistan from a United Nation's website. You can use any of these procedures – depending on the browser you are using: Google Chrome or Mozilla Firefox.

Scraping Data using Browser Extension for Google Chrome

Download Browser Extension to Scrape Data

1. Open your Chrome browser
2. To download **Scraper Extension**, go to https://chrome.google.com/webstore/detail/scraper/mbigbapnjcgaffohmbkdlecaccepngjd?hl=en
3. In the 'Scraper' window that opens, click the blue **+ ADD TO CHROME** button.



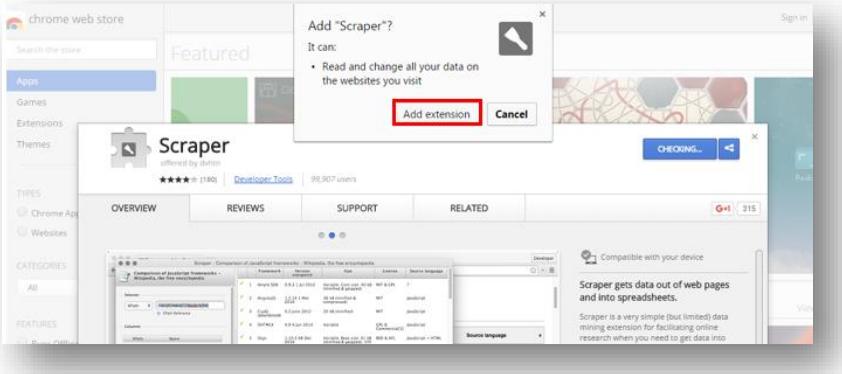
Scraper gets data out of web pages and into spreadsheets.

Scraper is a very simple (but limited) data mining extension for facilitating online research when you need to get data into spreadsheet form quickly. It is intended as an easy-to-use tool for intermediate to advanced users who are comfortable with XPath.

* 1.7
- feature: copy data to clipboard (as tab-separated values)
- fix: upgraded oauth for Google Docs

Additional Information
Version: 1.7
Updated: April 20, 2015
Size: 1.97MiB
Language: English

4. In the popup window that appears, click the **Add extension** button.



Add "Scraper"?
It can:
- Read and change all your data on the websites you visit

5. The extension is now added to chrome.

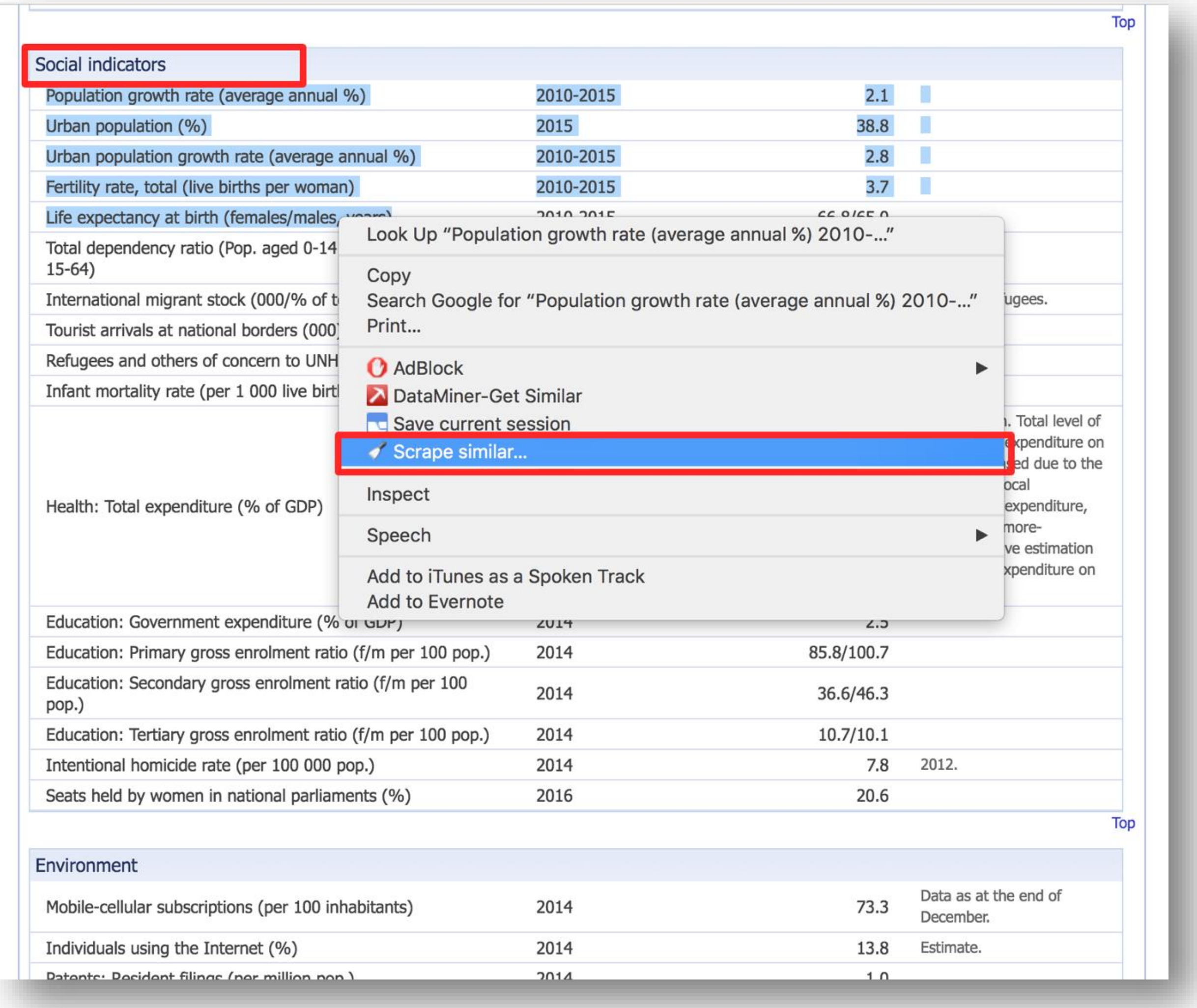
Scraping Data from a Website

6. Now, open a new Chrome window

7. Go to: http://data.un.org/CountryProfile.aspx?crName=Pakistan

8. Once the page opens, scroll down to reach the table with the heading "Social Indicators"

9. Now use the mouse pointer to highlight first three rows of the "Social Indicators" table, then right-click and select **Scrape similar**… (Note, if you select the entire table, the tool will not work. You have to select a portion of the table and Google will automatically find the rest of the table).

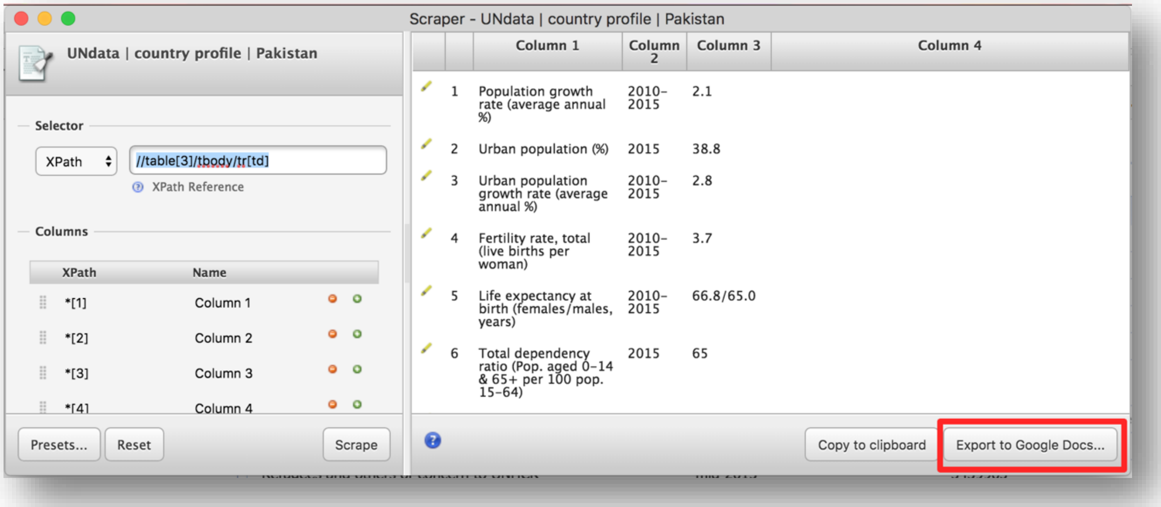


Top
Look Up "Population growth rate (average annual %) 2010-…"
Copy
Search Google for "Population growth rate (average annual %) 2010-…"
Print…
AdBlock ▶
DataMiner-Get Similar
Save current session
Scrape similar…
Inspect
Speech ▶
Add to iTunes as a Spoken Track
Add to Evernote
Top

Social indicators			
Population growth rate (average annual %)	2010-2015	2.1	
Urban population (%)	2015	38.8	
Urban population growth rate (average annual %)	2010-2015	2.8	
Fertility rate, total (live births per woman)	2010-2015	3.7	
Life expectancy at birth (females/males, years)	2010-2015	66.8/65.0	
Total dependency ratio (Pop. aged 0-14 15-64)			
International migrant stock (000/% of t			ugees.
Tourist arrivals at national borders (000)			
Refugees and others of concern to UNH			
Infant mortality rate (per 1 000 live birt			
Health: Total expenditure (% of GDP)			. Total level of xpenditure on ed due to the ocal expenditure, more- ve estimation xpenditure on
Education: Government expenditure (% of GDP)	2014	2.5	
Education: Primary gross enrolment ratio (f/m per 100 pop.)	2014	85.8/100.7	
Education: Secondary gross enrolment ratio (f/m per 100 pop.)	2014	36.6/46.3	
Education: Tertiary gross enrolment ratio (f/m per 100 pop.)	2014	10.7/10.1	
Intentional homicide rate (per 100 000 pop.)	2014	7.8	2012.
Seats held by women in national parliaments (%)	2016	20.6	

Environment			
Mobile-cellular subscriptions (per 100 inhabitants)	2014	73.3	Data as at the end of December.
Individuals using the Internet (%)	2014	13.8	Estimate.
Patents: Resident filings (per million pop.)	2014	1.0	

10. In the Scraper window that opens, click the **Export to Google Docs…** button.



Scraper - UNdata | country profile | Pakistan

UNdata | country profile | Pakistan

Selector

XPath //table[3]/tbody/tr[td]

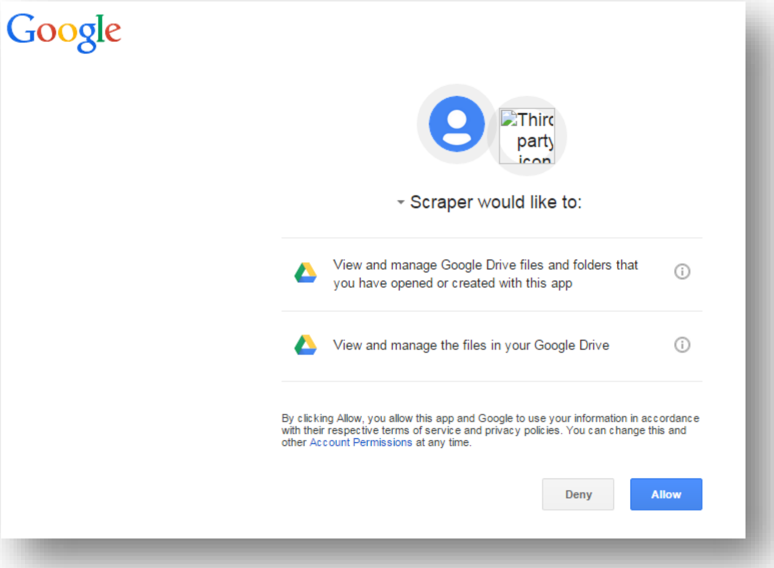
XPath Reference

Columns

XPath	Name
*[1]	Column 1
*[2]	Column 2
*[3]	Column 3
*[4]	Column 4

	Column 1	Column 2	Column 3	Column 4
1	Population growth rate (average annual %)	2010-2015	2.1	
2	Urban population (%)	2015	38.8	
3	Urban population growth rate (average annual %)	2010-2015	2.8	
4	Fertility rate, total (live births per woman)	2010-2015	3.7	
5	Life expectancy at birth (females/males, years)	2010-2015	66.8/65.0	
6	Total dependency ratio (Pop. aged 0-14 & 65+ per 100 pop. 15-64)	2015	65	

11. If you are not signed in to Google, Chrome asks you to sing-in using your Gmail id. A window opens seeking permission for the Scrpaer extension. Click **Allow**.



Google

Third party icon

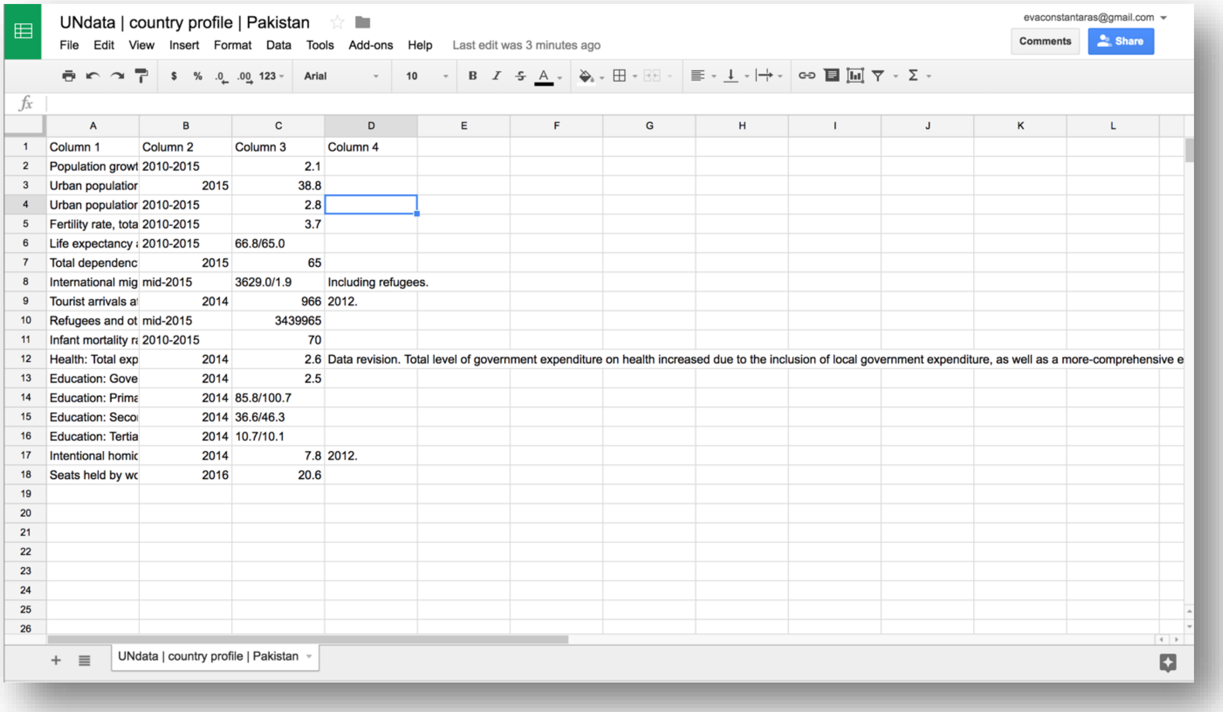
Scraper would like to:

View and manage Google Drive files and folders that you have opened or created with this app

View and manage the files in your Google Drive

By clicking Allow, you allow this app and Google to use your information in accordance with their respective terms of service and privacy policies. You can change this and other Account Permissions at any time.

12. A Google Spreadsheet now opens showing the Social Indicators data scraped from the UN website:


UNdata | country profile | Pakistan
File Edit View Insert Format Data Tools Add-ons Help
Last edit was 3 minutes ago
evaconstantaras@gmail.com
Comments
Share
Arial
10
UNdata | country profile | Pakistan

	A	B	C	D
1	Column 1	Column 2	Column 3	Column 4
2	Population growt	2010-2015	2.1	
3	Urban populatior	2015	38.8	
4	Urban populatior	2010-2015	2.8	
5	Fertility rate, tota	2010-2015	3.7	
6	Life expectancy a	2010-2015	66.8/65.0	
7	Total dependenc	2015	65	
8	International mig	mid-2015	3629.0/1.9	Including refugees.
9	Tourist arrivals a	2014	966	2012.
10	Refugees and ot	mid-2015	3439965	
11	Infant mortality ra	2010-2015	70	
12	Health: Total exp	2014	2.6	Data revision. Total level of government expenditure on health increased due to the inclusion of local government expenditure, as well as a more-comprehensive e
13	Education: Gove	2014	2.5	
14	Education: Prima	2014	85.8/100.7	
15	Education: Seco	2014	36.6/46.3	
16	Education: Tertia	2014	10.7/10.1	
17	Intentional homic	2014	7.8	2012.
18	Seats held by wc	2016	20.6	

Scraping Data using Browser Add-On for Mozilla Firefox

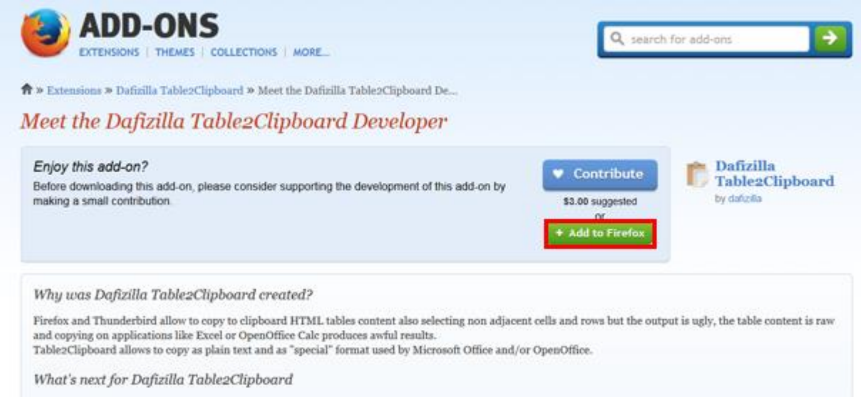
Download Browser Add-On to Scrape Data

13. Open Firefox browser.

14. To download an add-on called **Dafizilla Table2Clipboard**, go to: https://addons.mozilla.org/en-us/firefox/addon/dafizilla-table2clipboard/

15. On this webpage, click **Continue to Download**

16. On the next page that opens, click **+ Add to Firefox**


ADD-ONS
EXTENSIONS | THEMES | COLLECTIONS | MORE...
search for add-ons
» Extensions » Dafizilla Table2Clipboard » Meet the Dafizilla Table2Clipboard De...
Contribute
$3.00 suggested
or
+ Add to Firefox
Dafizilla Table2Clipboard
by dafizilla

Meet the Dafizilla Table2Clipboard Developer

Enjoy this add-on?

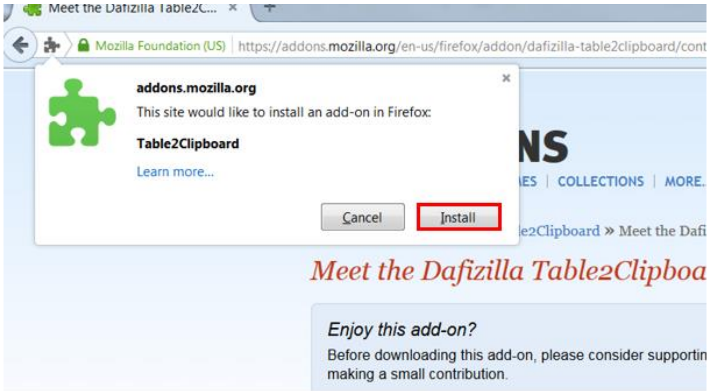
Before downloading this add-on, please consider supporting the development of this add-on by making a small contribution.

Why was Dafizilla Table2Clipboard created?

Firefox and Thunderbird allow to copy to clipboard HTML tables content also selecting non adjacent cells and rows but the output is ugly, the table content is raw and copying on applications like Excel or OpenOffice Calc produces awful results.
Table2Clipboard allows to copy as plain text and as "special" format used by Microsoft Office and/or OpenOffice.

What's next for Dafizilla Table2Clipboard

17. In the pop-up window that appears, click **Install**.


Meet the Dafizilla Table2C...
Mozilla Foundation (US)
https://addons.mozilla.org/en-us/firefox/addon/dafizilla-table2clipboard/cont
Cancel
Install
NS
ES | COLLECTIONS | MORE.
le2Clipboard » Meet the Dafi

addons.mozilla.org

This site would like to install an add-on in Firefox:

Table2Clipboard

Learn more...

Meet the Dafizilla Table2Clipboa

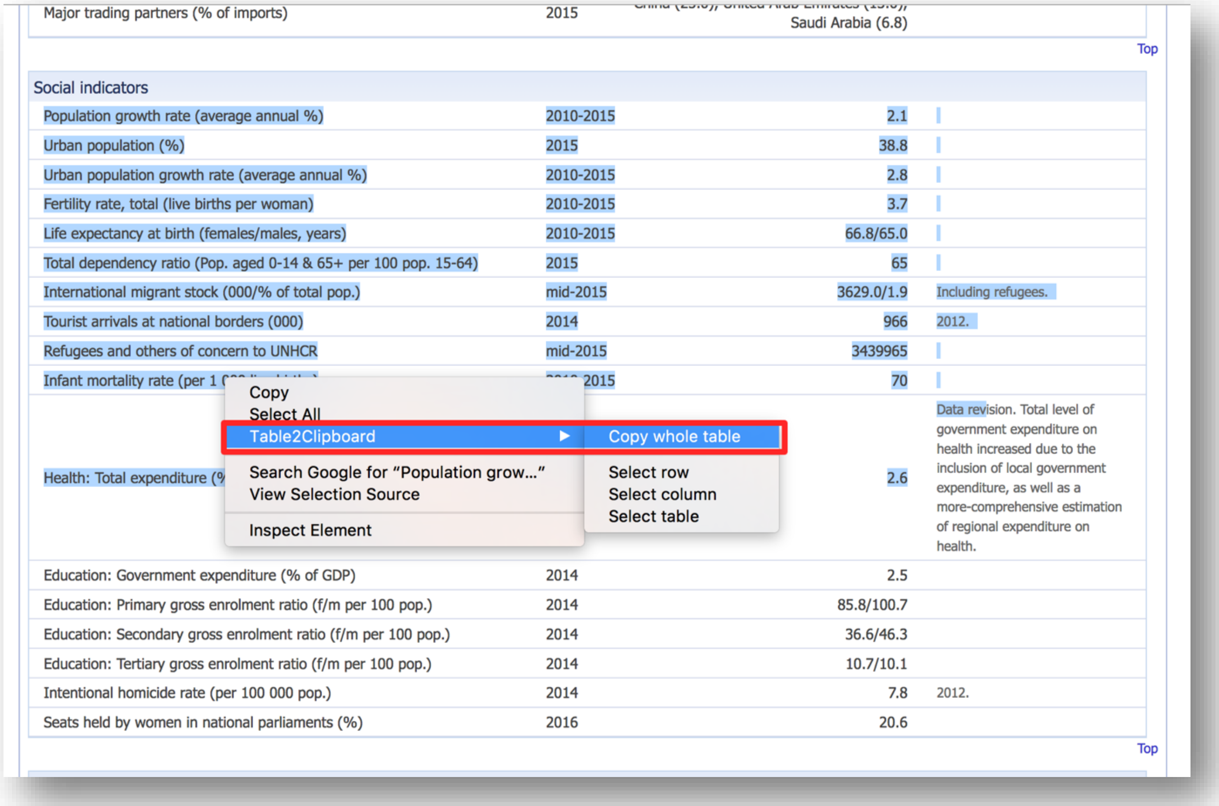
Enjoy this add-on?

Before downloading this add-on, please consider supportin making a small contribution.

18. In the next pop-up window, click **Restart Now**. Firefox closes and reopens – the add-on is now added to Firefox.

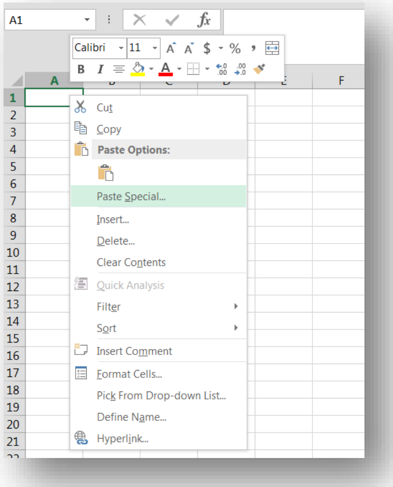
Scraping Data from a Website

19. Now, open a new Firefox window.
20. Go to: http://data.un.org/CountryProfile.aspx?crName=Pakistan
21. Once the page opens, scroll down to reach the table with the heading “Social Indicators”
22. Now use the mouse pointer to highlight first three rows of the “Social Indicators” table, then right-click and select T**able2Clipboard > Copy whole table**

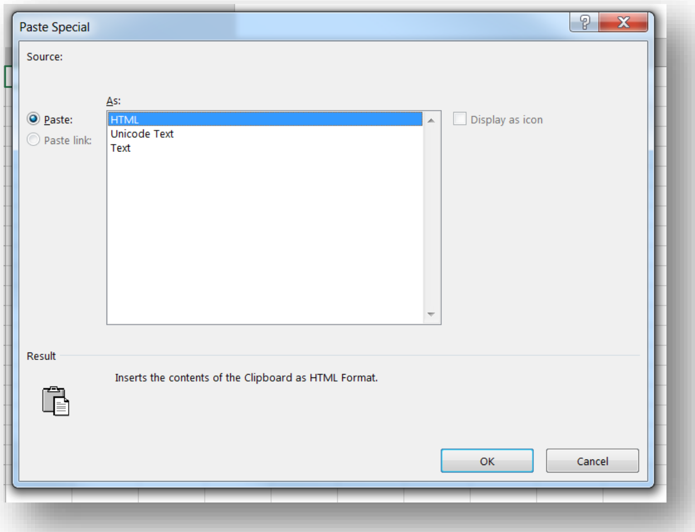


Major trading partners (% of imports)	2015	Saudi Arabia (6.8)	
Social indicators			
Population growth rate (average annual %)	2010-2015	2.1	
Urban population (%)	2015	38.8	
Urban population growth rate (average annual %)	2010-2015	2.8	
Fertility rate, total (live births per woman)	2010-2015	3.7	
Life expectancy at birth (females/males, years)	2010-2015	66.8/65.0	
Total dependency ratio (Pop. aged 0-14 & 65+ per 100 pop. 15-64)	2015	65	
International migrant stock (000/% of total pop.)	mid-2015	3629.0/1.9	Including refugees.
Tourist arrivals at national borders (000)	2014	966	2012.
Refugees and others of concern to UNHCR	mid-2015	3439965	
Infant mortality rate (per 1	2015	70	
Health: Total expenditure (%		2.6	Data revision. Total level of government expenditure on health increased due to the inclusion of local government expenditure, as well as a more-comprehensive estimation of regional expenditure on health.
Education: Government expenditure (% of GDP)	2014	2.5	
Education: Primary gross enrolment ratio (f/m per 100 pop.)	2014	85.8/100.7	
Education: Secondary gross enrolment ratio (f/m per 100 pop.)	2014	36.6/46.3	
Education: Tertiary gross enrolment ratio (f/m per 100 pop.)	2014	10.7/10.1	
Intentional homicide rate (per 100 000 pop.)	2014	7.8	2012.
Seats held by women in national parliaments (%)	2016	20.6	

23. Now, open a new Microsoft Excel file.
24. In the cell A1 of the spreadsheet, right-click and select **Paste Special**



25. In the ‘Paste Special’ window, check if **HTML** is selected, and then click **OK.**



Source:

● Paste:
○ Paste link:

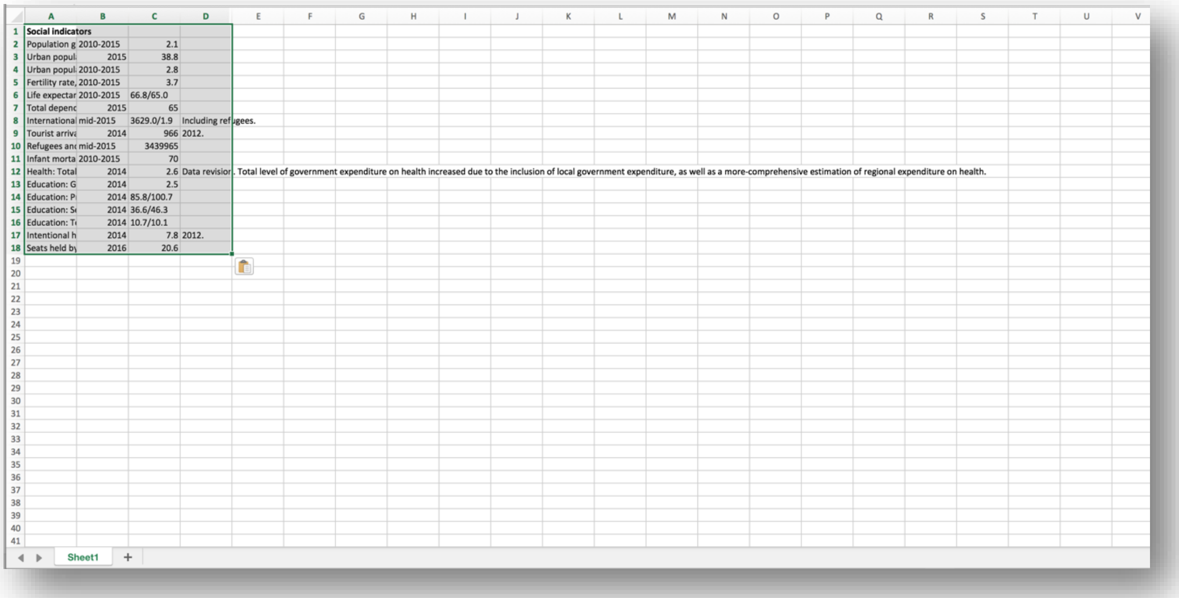
As:
HTML
Unicode Text
Text

□ Display as icon

Result

Inserts the contents of the Clipboard as HTML Format.

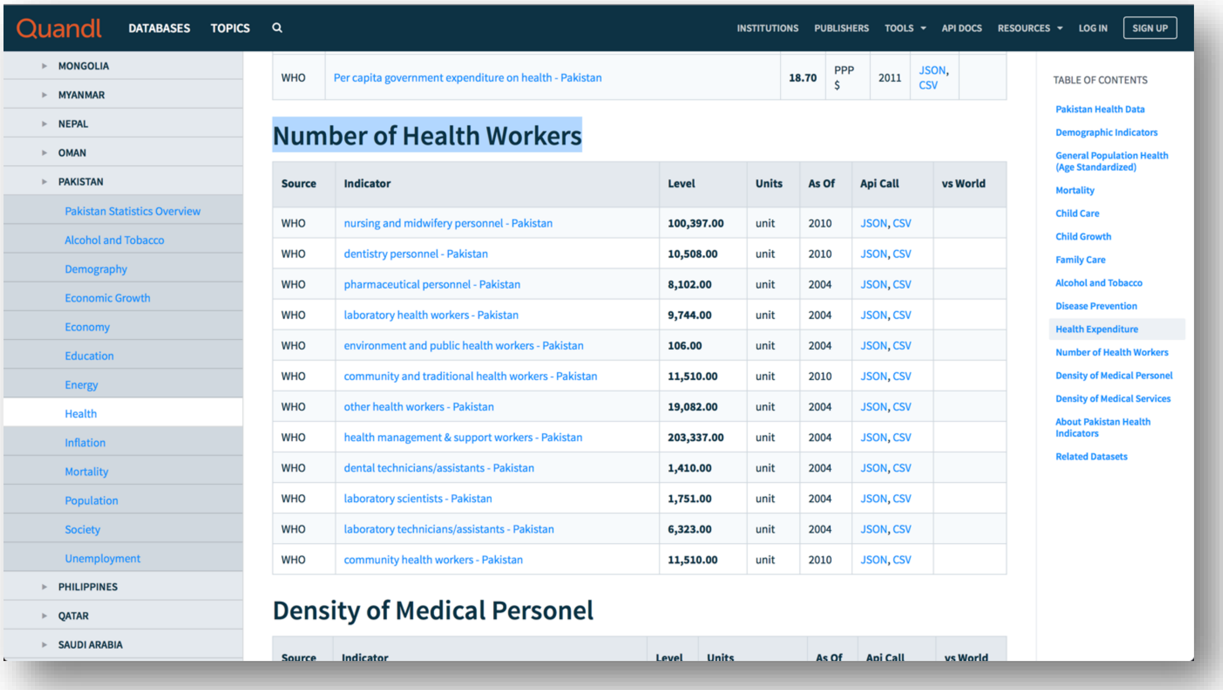
26. The social indicators data table is now available in Excel.



	A	B	C	D
1	**Social indicators**			
2	Population g	2010-2015	2.1	
3	Urban popul	2015	38.8	
4	Urban popul	2010-2015	2.8	
5	Fertility rate,	2010-2015	3.7	
6	Life expectar	2010-2015	66.8/65.0	
7	Total depenc	2015	65	
8	International	mid-2015	3629.0/1.9	Including refugees.
9	Tourist arriva	2014	966	2012.
10	Refugees anc	mid-2015	3439965	
11	Infant morta	2010-2015	70	
12	Health: Total	2014	2.6	Data revision. Total level of government expenditure on health increased due to the inclusion of local government expenditure, as well as a more-comprehensive estimation of regional expenditure on health.
13	Education: G	2014	2.5	
14	Education: P	2014	85.8/100.7	
15	Education: S	2014	36.6/46.3	
16	Education: T	2014	10.7/10.1	
17	Intentional h	2014	7.8	2012.
18	Seats held by	2016	20.6	

Exercise: Scrape Data Using Browser Extensions

Use the Chrome or Firefox browser to scrape data from the ***Number of Health Workers*** table available on this web page: https://www.quandl.com/collections/pakistan/pakistan-health-data



WHO	Per capita government expenditure on health - Pakistan	18.70	PPP $	2011	JSON, CSV	

Number of Health Workers

Source	Indicator	Level	Units	As Of	Api Call	vs World
WHO	nursing and midwifery personnel - Pakistan	**100,397.00**	unit	2010	JSON, CSV	
WHO	dentistry personnel - Pakistan	**10,508.00**	unit	2010	JSON, CSV	
WHO	pharmaceutical personnel - Pakistan	**8,102.00**	unit	2004	JSON, CSV	
WHO	laboratory health workers - Pakistan	**9,744.00**	unit	2004	JSON, CSV	
WHO	environment and public health workers - Pakistan	**106.00**	unit	2004	JSON, CSV	
WHO	community and traditional health workers - Pakistan	**11,510.00**	unit	2010	JSON, CSV	
WHO	other health workers - Pakistan	**19,082.00**	unit	2004	JSON, CSV	
WHO	health management & support workers - Pakistan	**203,337.00**	unit	2004	JSON, CSV	
WHO	dental technicians/assistants - Pakistan	**1,410.00**	unit	2004	JSON, CSV	
WHO	laboratory scientists - Pakistan	**1,751.00**	unit	2004	JSON, CSV	
WHO	laboratory technicians/assistants - Pakistan	**6,323.00**	unit	2004	JSON, CSV	
WHO	community health workers - Pakistan	**11,510.00**	unit	2010	JSON, CSV	

Density of Medical Personel

Source	Indicator	Level	Units	As Of	Api Call	vs World

27. Scrape the data from Number of Health Workers using chrome Scraper Tool or Firefox Dafizilla Table2Clipboard.

28. Chose another table from https://www.quandl.com/collections/pakistan/pakistan-demography-data and scrape it.

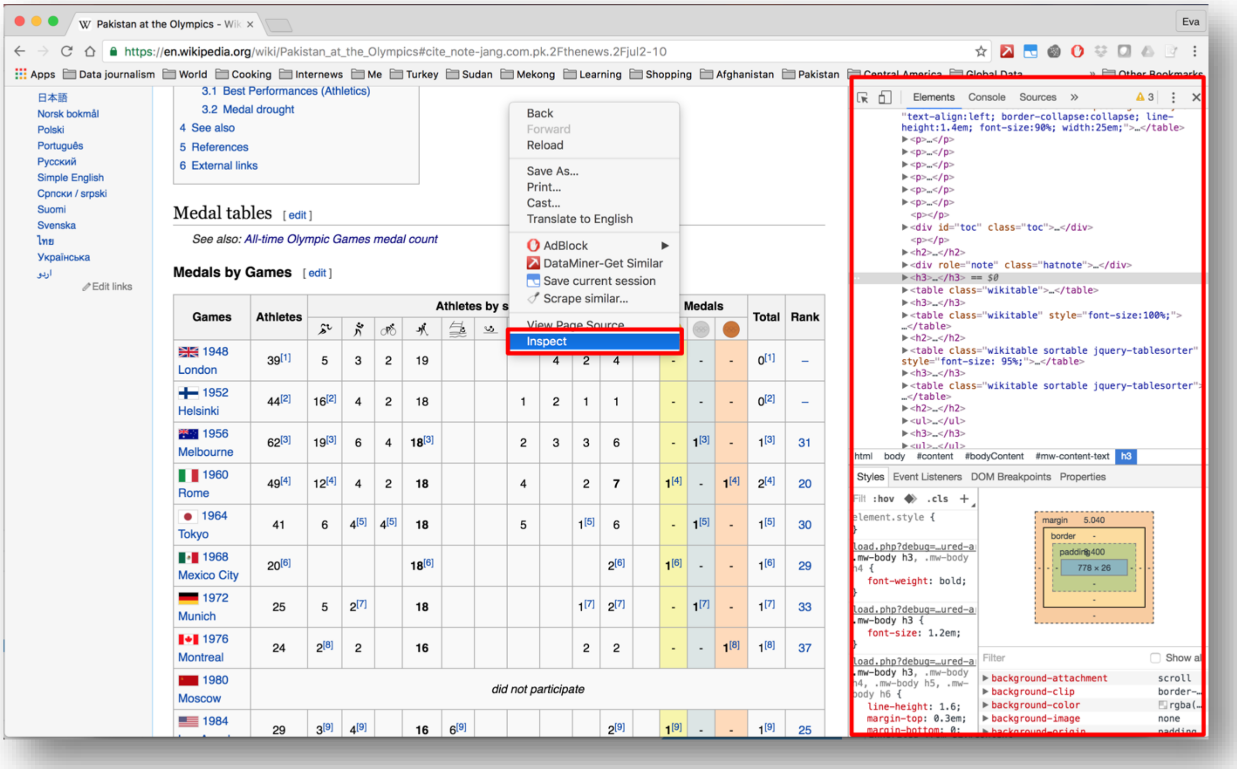
Task 3: Scraping Data Using ImportHTML

Code Source from a webpage

Behind of all text, table, photos, videos that a web page contain there is sourcecode. This code determines what is on the page, how it looks and what it does.

29. Open [https://en.wikipedia.org/wiki/Pakistan_at_the_Olympics - cite_note-jang.com.pk.2Fthenews.2Fjul2-10](https://en.wikipedia.org/wiki/Pakistan_at_the_Olympics#cite_note-jang.com.pk.2Fthenews.2Fjul2-10) either in google chrome or firefox

30. To see the source code right click and click in Inspect (Google Chrome) or Inspect Element (Firefox)



Medal tables [edit]

See also: All-time Olympic Games medal count

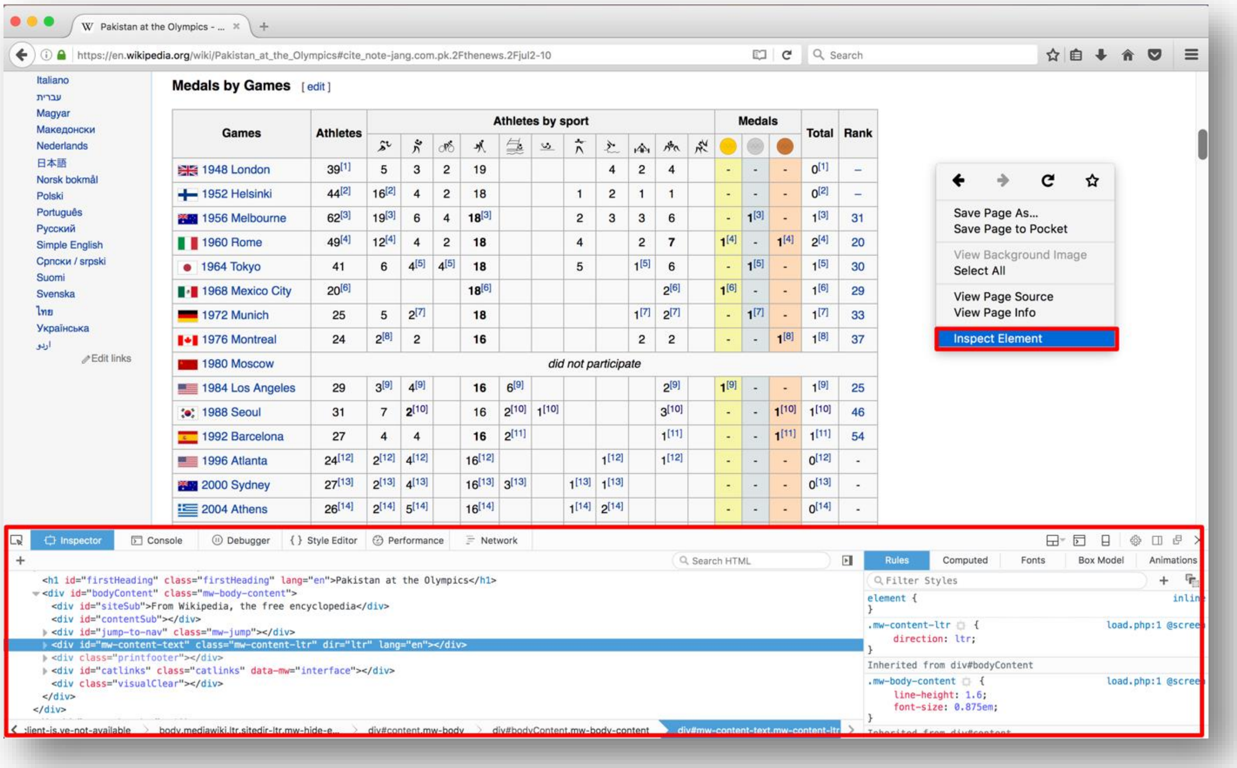
Medals by Games [edit]

Games	Athletes	Athletes by s											Medals			Total	Rank
1948 London	39[1]	5	3	2	19				4	2	4		-	-	-	0[1]	–
1952 Helsinki	44[2]	16[2]	4	2	18			1	2	1	1		-	-	-	0[2]	–
1956 Melbourne	62[3]	19[3]	6	4	18[3]			2	3	3	6		-	1[3]	-	1[3]	31
1960 Rome	49[4]	12[4]	4	2	18			4		2	7		1[4]	-	1[4]	2[4]	20
1964 Tokyo	41	6	4[5]	4[5]	18			5		1[5]	6		-	1[5]	-	1[5]	30
1968 Mexico City	20[6]				18[6]						2[6]		1[6]	-	-	1[6]	29
1972 Munich	25	5	2[7]		18					1[7]	2[7]		-	1[7]	-	1[7]	33
1976 Montreal	24	2[8]	2		16					2	2		-	-	1[8]	1[8]	37
1980 Moscow	*did not participate*																
1984	29	3[9]	4[9]		16	6[9]					2[9]		1[9]	-	-	1[9]	25

```
"text-align:left; border-collapse:collapse; line-
height:1.4em; font-size:90%; width:25em;">…</table>
▶<p>…</p>
▶<p>…</p>
▶<p>…</p>
▶<p>…</p>
▶<p>…</p>
▶<p>…</p>
 <p></p>
▶<div id="toc" class="toc">…</div>
 <p></p>
▶<h2>…</h2>
▶<div role="note" class="hatnote">…</div>
▶<h3>…</h3> == $0
▶<table class="wikitable">…</table>
▶<h3>…</h3>
▶<table class="wikitable" style="font-size:100%;">
…</table>
▶<h2>…</h2>
▶<table class="wikitable sortable jquery-tablesorter"
style="font-size: 95%;">…</table>
▶<h3>…</h3>
▶<table class="wikitable sortable jquery-tablesorter">
…</table>
▶<h2>…</h2>
▶<ul>…</ul>
▶<h3>…</h3>
▶<ul>…</ul>
```

```
element.style {
}
load.php?debug=…ured-a
.mw-body h3, .mw-body
h4 {
  font-weight: bold;
}
load.php?debug=…ured-a
.mw-body h3 {
  font-size: 1.2em;
}
load.php?debug=…ured-a
.mw-body h3, .mw-body
h4, .mw-body h5, .mw-
body h6 {
  line-height: 1.6;
  margin-top: 0.3em;
  margin-bottom: 0;
```

```
▶ background-attachment     scroll
▶ background-clip           border-…
▶ background-color          rgba(…
▶ background-image          none
▶ background-origin         padding
```



Medals by Games [edit]

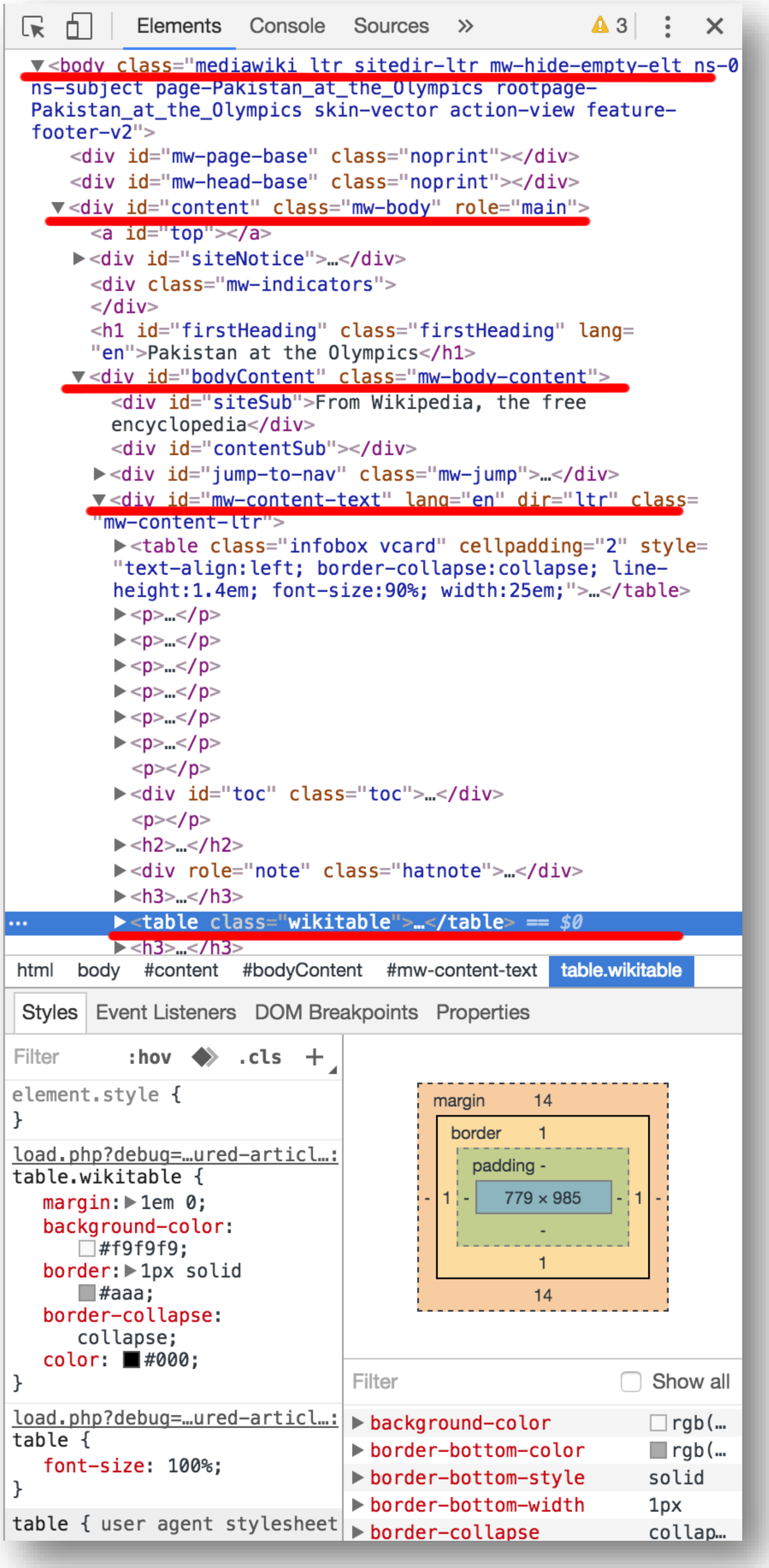
Games	Athletes	Athletes by sport											Medals			Total	Rank
													●	●	●		
1948 London	39[1]	5	3	2	19				4	2	4		-	-	-	0[1]	–
1952 Helsinki	44[2]	16[2]	4	2	18			1	2	1	1		-	-	-	0[2]	–
1956 Melbourne	62[3]	19[3]	6	4	18[3]			2	3	3	6		-	1[3]	-	1[3]	31
1960 Rome	49[4]	12[4]	4	2	18			4		2	7		1[4]	-	1[4]	2[4]	20
1964 Tokyo	41	6	4[5]	4[5]	18			5		1[5]	6		-	1[5]	-	1[5]	30
1968 Mexico City	20[6]				18[6]						2[6]		1[6]	-	-	1[6]	29
1972 Munich	25	5	2[7]		18					1[7]	2[7]		-	1[7]	-	1[7]	33
1976 Montreal	24	2[8]	2		16					2	2		-	-	1[8]	1[8]	37
1980 Moscow	did not participate																
1984 Los Angeles	29	3[9]	4[9]		16	6[9]					2[9]		1[9]	-	-	1[9]	25
1988 Seoul	31	7	2[10]		16	2[10]	1[10]				3[10]		-	-	1[10]	1[10]	46
1992 Barcelona	27	4	4		16	2[11]					1[11]		-	-	1[11]	1[11]	54
1996 Atlanta	24[12]	2[12]	4[12]		16[12]				1[12]		1[12]		-	-	-	0[12]	-
2000 Sydney	27[13]	2[13]	4[13]		16[13]	3[13]		1[13]	1[13]				-	-	-	0[13]	-
2004 Athens	26[14]	2[14]	5[14]		16[14]			1[14]	2[14]				-	-	-	0[14]	-

```
  <h1 id="firstHeading" class="firstHeading" lang="en">Pakistan at the Olympics</h1>
 ▾<div id="bodyContent" class="mw-body-content">
    <div id="siteSub">From Wikipedia, the free encyclopedia</div>
    <div id="contentSub"></div>
  ▸ <div id="jump-to-nav" class="mw-jump"></div>
  ▸ <div id="mw-content-text" class="mw-content-ltr" dir="ltr" lang="en"></div>
  ▸ <div class="printfooter"></div>
  ▸ <div id="catlinks" class="catlinks" data-mw="interface"></div>
    <div class="visualClear"></div>
  </div>
</div>
```

```
element {                                  inlin
}
.mw-content-ltr {                 load.php:1 @scree
    direction: ltr;
}
Inherited from div#bodyContent
.mw-body-content {                load.php:1 @scree
    line-height: 1.6;
    font-size: 0.875em;
}
```

:lient-js.ve-not-available
body.mediawiki.ltr.sitedir-ltr.mw-hide-e...
div#content.mw-body
div#bodyContent.mw-body-content
div#mw-content-text.mw-content-ltr

31. When you move the cursor inside the code the parts in the webpage will be selected. Inside the source code go to body class - content – mw-contect-text – wiketable

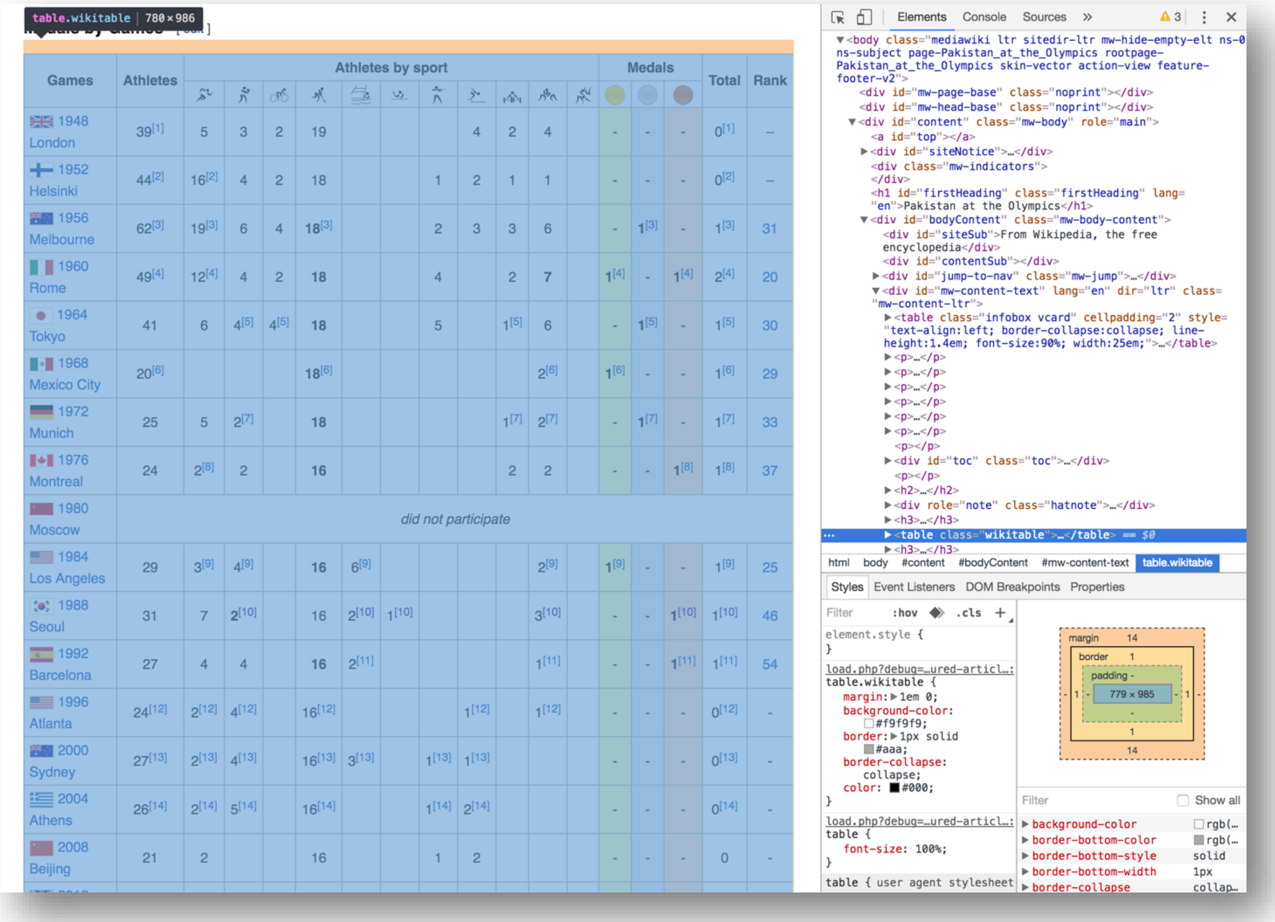


```
▼<body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0
ns-subject page-Pakistan_at_the_Olympics rootpage-
Pakistan_at_the_Olympics skin-vector action-view feature-
footer-v2">
    <div id="mw-page-base" class="noprint"></div>
    <div id="mw-head-base" class="noprint"></div>
  ▼<div id="content" class="mw-body" role="main">
      <a id="top"></a>
    ►<div id="siteNotice">…</div>
      <div class="mw-indicators">
      </div>
      <h1 id="firstHeading" class="firstHeading" lang=
      "en">Pakistan at the Olympics</h1>
    ▼<div id="bodyContent" class="mw-body-content">
        <div id="siteSub">From Wikipedia, the free
        encyclopedia</div>
        <div id="contentSub"></div>
      ►<div id="jump-to-nav" class="mw-jump">…</div>
      ▼<div id="mw-content-text" lang="en" dir="ltr" class=
       "mw-content-ltr">
        ►<table class="infobox vcard" cellpadding="2" style=
         "text-align:left; border-collapse:collapse; line-
         height:1.4em; font-size:90%; width:25em;">…</table>
        ►<p>…</p>
        ►<p>…</p>
        ►<p>…</p>
        ►<p>…</p>
        ►<p>…</p>
        ►<p>…</p>
          <p></p>
        ►<div id="toc" class="toc">…</div>
          <p></p>
        ►<h2>…</h2>
        ►<div role="note" class="hatnote">…</div>
        ►<h3>…</h3>
…       ►<table class="wikitable">…</table> == $0
        ►<h3>…</h3>
```

```
element.style {
}
load.php?debug=…ured-articl…:
table.wikitable {
   margin:►1em 0;
   background-color:
      □#f9f9f9;
   border:►1px solid
      ■#aaa;
   border-collapse:
      collapse;
   color: ■#000;
}
load.php?debug=…ured-articl…:
table {
   font-size: 100%;
}
table { user agent stylesheet
```

```
► background-color        □rgb(…
► border-bottom-color     ■rgb(…
► border-bottom-style     solid
► border-bottom-width     1px
► border-collapse         collap…
```

32. You will see the Table selected. As you can see every part in the table is selected as you move the cursor in the code, links, pictures, and data.



Games	Athletes	Athletes by sport											Medals			Total	Rank
1948 London	39[1]	5	3	2	19				4	2	4		-	-	-	0[1]	–
1952 Helsinki	44[2]	16[2]	4	2	18			1	2	1	1		-	-	-	0[2]	–
1956 Melbourne	62[3]	19[3]	6	4	18[3]			2	3	3	6		-	1[3]	-	1[3]	31
1960 Rome	49[4]	12[4]	4	2	18			4		2	7		1[4]	-	1[4]	2[4]	20
1964 Tokyo	41	6	4[5]	4[5]	18			5		1[5]	6		-	1[5]	-	1[5]	30
1968 Mexico City	20[6]				18[6]						2[6]		1[6]	-	-	1[6]	29
1972 Munich	25	5	2[7]		18					1[7]	2[7]		-	1[7]	-	1[7]	33
1976 Montreal	24	2[8]	2		16					2	2		-	-	1[8]	1[8]	37
1980 Moscow	*did not participate*																
1984 Los Angeles	29	3[9]	4[9]		16	6[9]					2[9]		1[9]	-	-	1[9]	25
1988 Seoul	31	7	2[10]		16	2[10]	1[10]				3[10]		-	-	1[10]	1[10]	46
1992 Barcelona	27	4	4		16	2[11]					1[11]		-	-	1[11]	1[11]	54
1996 Atlanta	24[12]	2[12]	4[12]		16[12]				1[12]		1[12]		-	-	-	0[12]	-
2000 Sydney	27[13]	2[13]	4[13]		16[13]	3[13]		1[13]	1[13]				-	-	-	0[13]	-
2004 Athens	26[14]	2[14]	5[14]		16[14]			1[14]	2[14]				-	-	-	0[14]	-
2008 Beijing	21	2			16			1	2				-	-	-	0	-

```
▼<body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0
ns-subject page-Pakistan_at_the_Olympics rootpage-
Pakistan_at_the_Olympics skin-vector action-view feature-
footer-v2">
    <div id="mw-page-base" class="noprint"></div>
    <div id="mw-head-base" class="noprint"></div>
  ▼<div id="content" class="mw-body" role="main">
      <a id="top"></a>
    ▶<div id="siteNotice">…</div>
      <div class="mw-indicators">
      </div>
      <h1 id="firstHeading" class="firstHeading" lang=
      "en">Pakistan at the Olympics</h1>
    ▼<div id="bodyContent" class="mw-body-content">
        <div id="siteSub">From Wikipedia, the free
        encyclopedia</div>
        <div id="contentSub"></div>
      ▶<div id="jump-to-nav" class="mw-jump">…</div>
      ▼<div id="mw-content-text" lang="en" dir="ltr" class=
      "mw-content-ltr">
        ▶<table class="infobox vcard" cellpadding="2" style=
        "text-align:left; border-collapse:collapse; line-
        height:1.4em; font-size:90%; width:25em;">…</table>
        ▶<p>…</p>
        ▶<p>…</p>
        ▶<p>…</p>
        ▶<p>…</p>
        ▶<p>…</p>
        ▶<p>…</p>
          <p></p>
        ▶<div id="toc" class="toc">…</div>
          <p></p>
        ▶<h2>…</h2>
        ▶<div role="note" class="hatnote">…</div>
        ▶<h3>…</h3>
...     ▶<table class="wikitable">…</table> == $0
        ▶<h3>…</h3>
```

```
element.style {
}
load.php?debug=…ured-articl…:
table.wikitable {
   margin:▶1em 0;
   background-color:
      □#f9f9f9;
   border:▶1px solid
      ■#aaa;
   border-collapse:
      collapse;
   color: ■#000;
}
load.php?debug=…ured-articl…:
table {
   font-size: 100%;
}
table { user agent stylesheet
```

```
▶ background-color       □rgb(…
▶ border-bottom-color    ■rgb(…
▶ border-bottom-style    solid
▶ border-bottom-width    1px
▶ border-collapse        collap…
```

Using ImportHTML

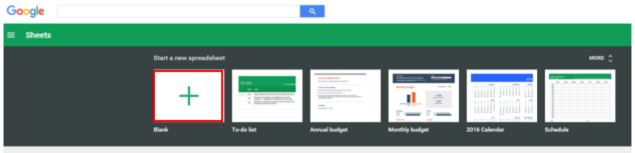
The Google Spreadsheet function **=importHTML(“”,“table”,N)** helps scrape a table from an HTML web page into a Google spreadsheet. Within this function:

- The URL of the target web page, and the target table element both need to be in double quotes.
- The number **N** identifies the table in the page (counting starts at 0) as the target table for data scraping.

Here is an example. Let’s use Google Spreadsheets to download data about the medals won by Pakistan in the Olympics:

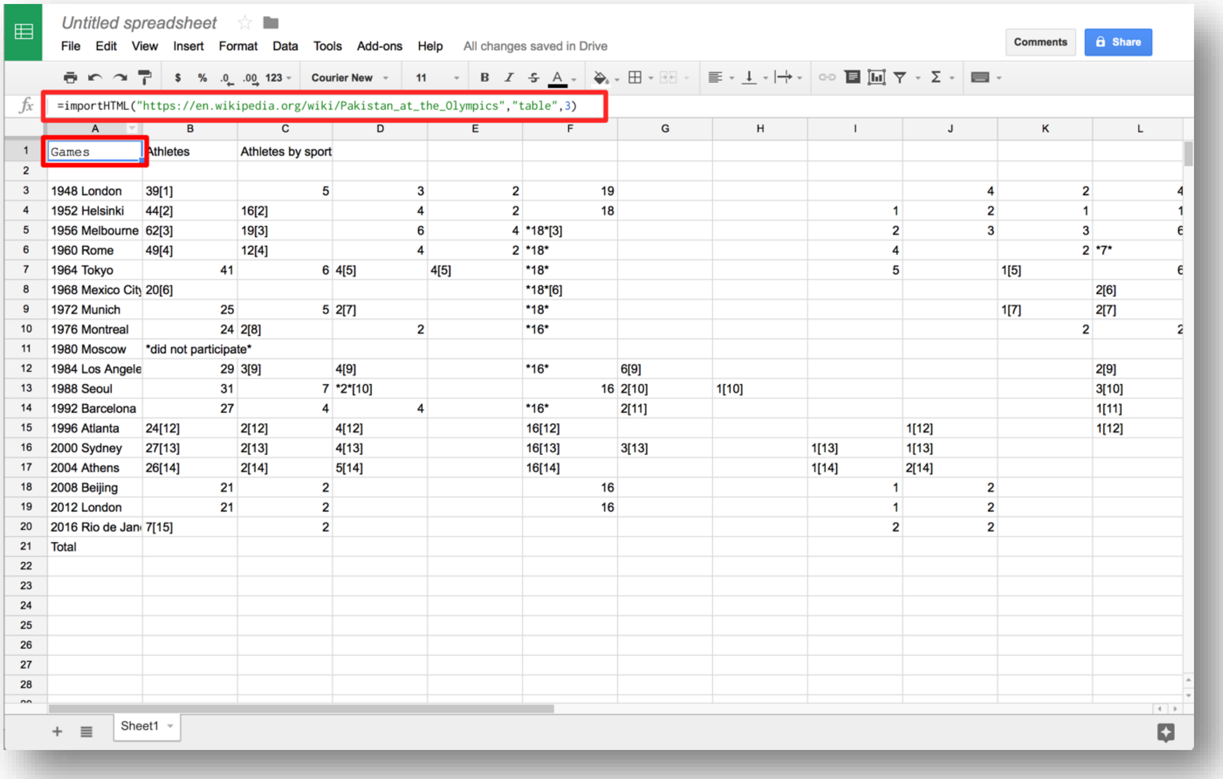
33. Open https://en.wikipedia.org/wiki/Pakistan_at_the_Olympics#cite_note-jang.com.pk.2Fthenews.2Fjul2-10 Scroll down the page to find “Medal by Games” table.

34. Now, open a blank Google Spreadsheet from: https://docs.google.com/spreadsheets/ (if you are not signed-in, Google will ask you to sign in).



35. This is the format for formulas in Google sheets: =function("url", "object" , number)

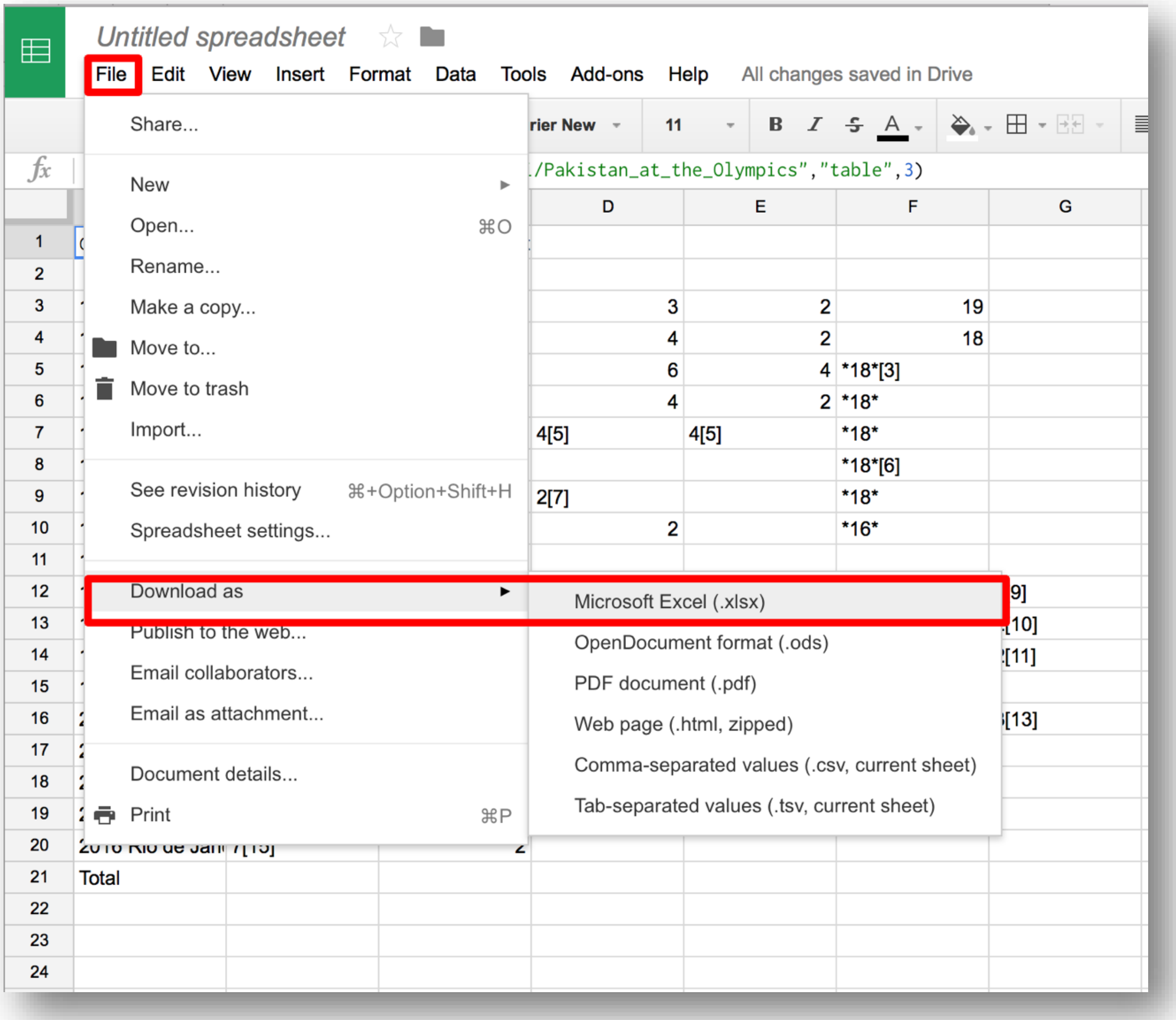
36. In the new Google Spreadsheet that opens, type the following function: **=importHTML("https://en.wikipedia.org/wiki/Pakistan_at_the_Olympics#cite_note-jang.com.pk.2Fthenews.2Fjul2-10","table",3)**



```
=importHTML("https://en.wikipedia.org/wiki/Pakistan_at_the_Olympics","table",3)
```

	A	B	C	D	E	F	G	H	I	J	K	L
1	Games	Athletes	Athletes by sport									
2												
3	1948 London	39[1]	5	3	2	19				4	2	4
4	1952 Helsinki	44[2]	16[2]	4	2	18			1	2	1	1
5	1956 Melbourne	62[3]	19[3]	6	4	*18*[3]			2	3	3	6
6	1960 Rome	49[4]	12[4]	4	2	*18*			4		2	*7*
7	1964 Tokyo	41	6	4[5]	4[5]	*18*			5		1[5]	6
8	1968 Mexico City	20[6]				*18*[6]						2[6]
9	1972 Munich	25	5	2[7]		*18*					1[7]	2[7]
10	1976 Montreal	24	2[8]	2		*16*					2	2
11	1980 Moscow	*did not participate*										
12	1984 Los Angele	29	3[9]	4[9]		*16*	6[9]					2[9]
13	1988 Seoul	31	7	*2*[10]		16	2[10]	1[10]				3[10]
14	1992 Barcelona	27	4	4		*16*	2[11]					1[11]
15	1996 Atlanta	24[12]	2[12]	4[12]		16[12]				1[12]		1[12]
16	2000 Sydney	27[13]	2[13]	4[13]		16[13]	3[13]		1[13]	1[13]		
17	2004 Athens	26[14]	2[14]	5[14]		16[14]			1[14]	2[14]		
18	2008 Beijing	21	2			16			1	2		
19	2012 London	21	2			16			1	2		
20	2016 Rio de Jan	7[15]	2						2	2		
21	Total											
22												
23												
24												
25												
26												
27												
28												

37. If you prefer to use Excel, click on File- Download as-Microsoft Excel.



Exercise

Scrape http://privatisation.gov.pk/?page_id=125

Complex Scraping: References

If you would want to learn to about more complex scraping applications, refer to these resources:

- **Import.io** is a browser based web scraping tool: https://www.import.io/ By following their easy step-by-step method you select the data you want to scrape and the tool does the rest.
- **Outwit Hub** is a software package that you can use on your PC or laptop: http://www.outwit.com/products/hub/ Note that the free version can only scrape 100 rows of data.
- **Regular Expressions** is a coding language that looks for a particular pattern of words, digits, or characters that recurs and extracts data using code: http://www.regular-expressions.info/
- **Morph.io** offers pre-written, editable scripts that use code not only to scrape but also to schedule and reformat scraping with reusable libraries: https://morph.io/