Data Fundamentals Lab: Scraping Data from the Web

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

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

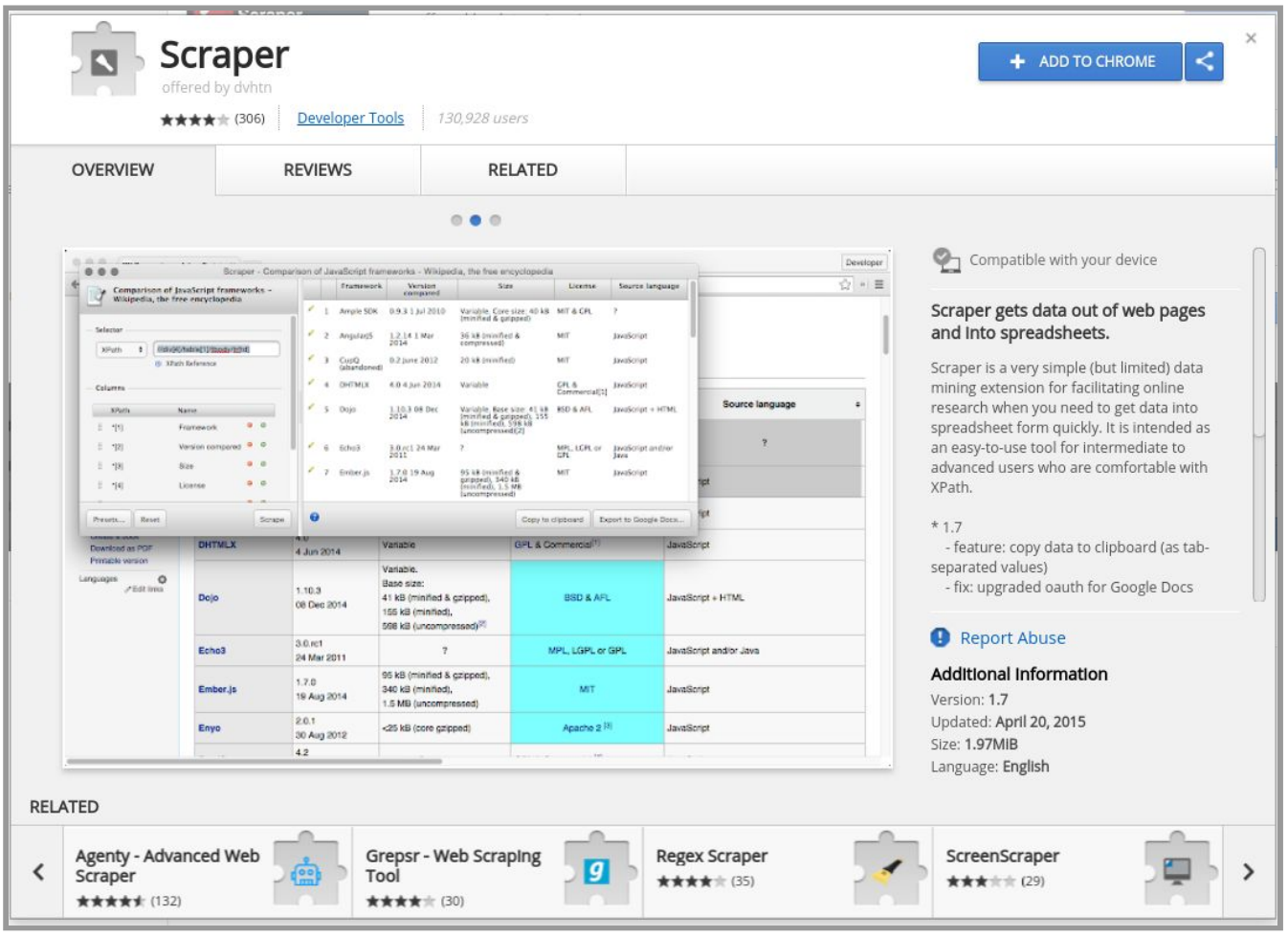
Here is an example, let's use a browser extension to scrape data about Tanzania from a United Nation's website. You can use any of these procedures – depending on the browser you are using: Google Chrome or Mozilla Firefox.

Scraping Data using Browser Extension for Google Chrome

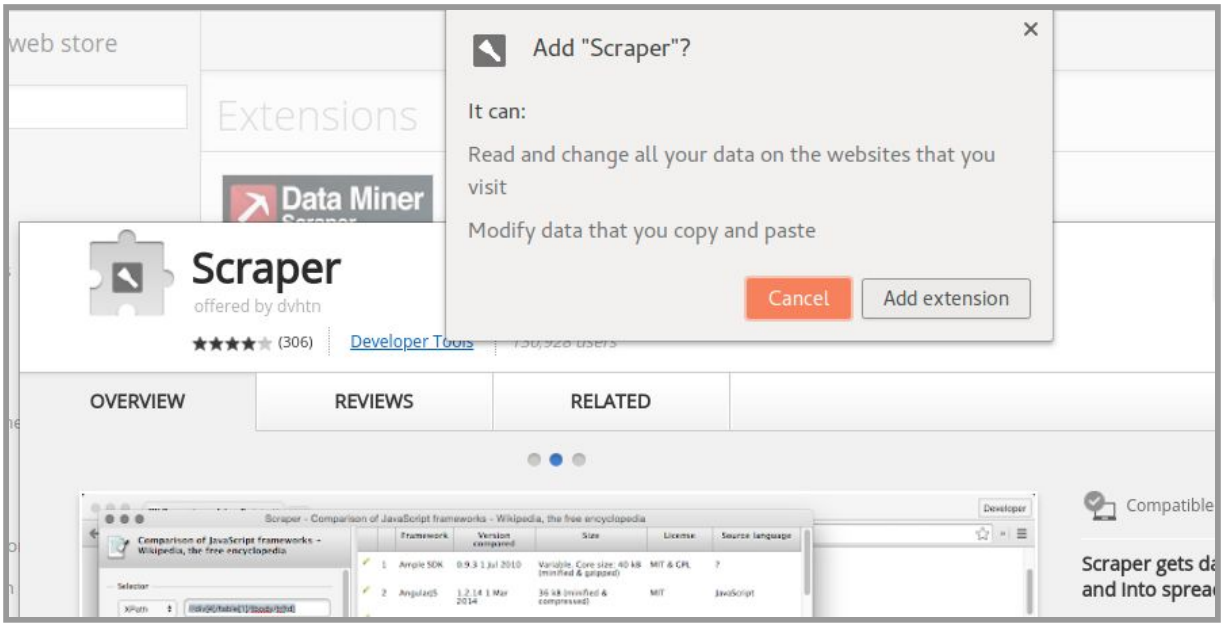
1. Open your Google Chrome Browser.

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.



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



5. The extension is now added to Chrome.

Scraping Data from a Website

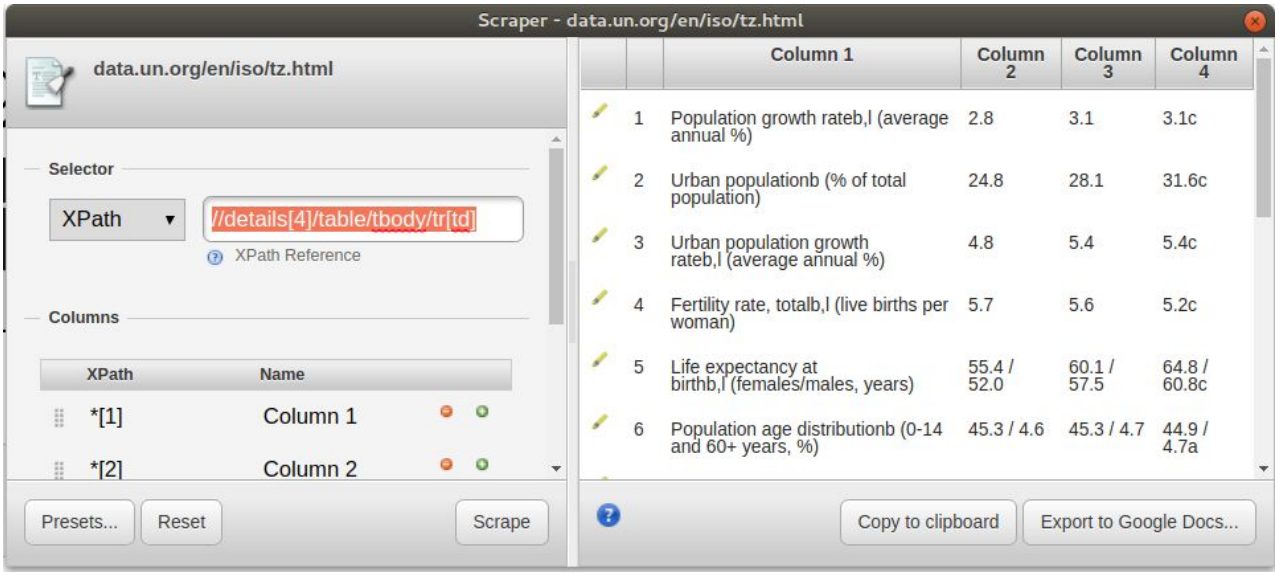
6. Now, open a new Chrome window.

7. Go to: http://data.un.org/en/iso/tz.html

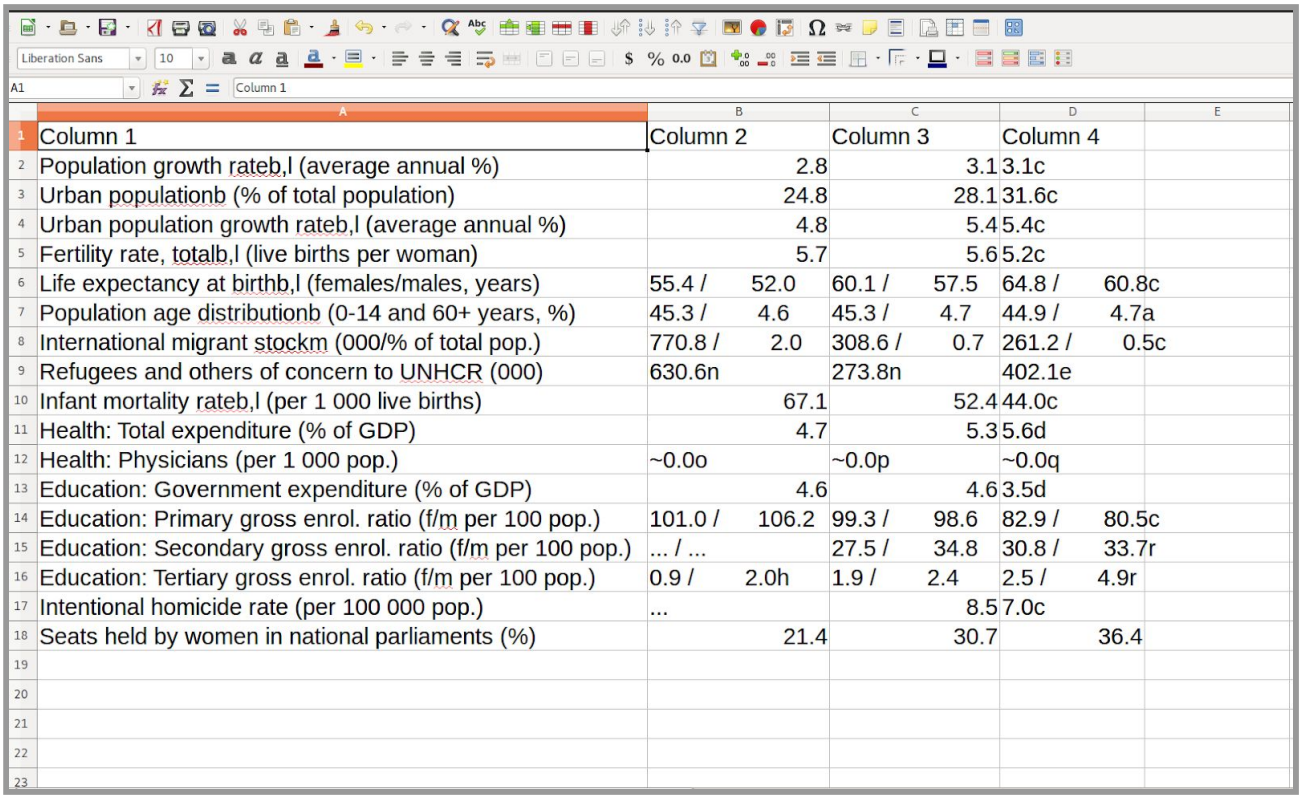
8. Once the page opens, click on the at the **"Social Indicators"** section to show the table.

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).

10. In the Scraper window that opens, click the **Copy to clipboard** button.



11. Open up a new sheet in your spreadsheet software of choice such as Microsoft Excel or LibreOffice Calc and paste the copied content into cell A1.

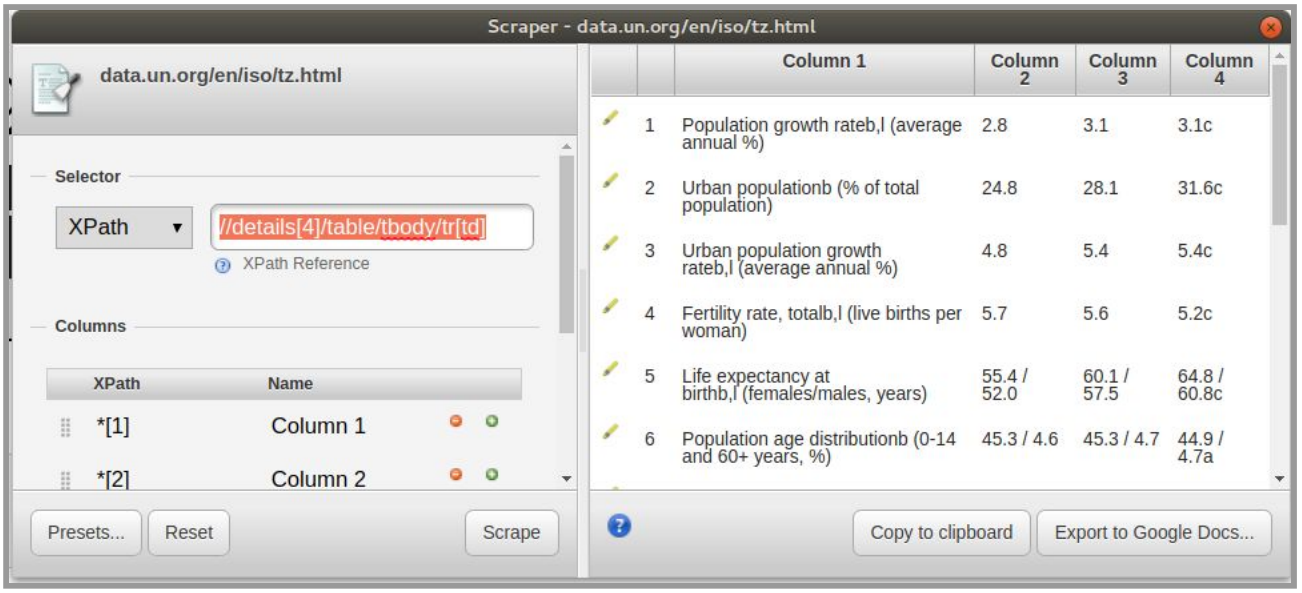


Column 1	Column 2	Column 3	Column 4
Population growth rateb,l (average annual %)	2.8	3.1	3.1c
Urban populationb (% of total population)	24.8	28.1	31.6c
Urban population growth rateb,l (average annual %)	4.8	5.4	5.4c
Fertility rate, totalb,l (live births per woman)	5.7	5.6	5.2c
Life expectancy at birthb,l (females/males, years)	55.4 / 52.0	60.1 / 57.5	64.8 / 60.8c
Population age distributionb (0-14 and 60+ years, %)	45.3 / 4.6	45.3 / 4.7	44.9 / 4.7a
International migrant stockm (000/% of total pop.)	770.8 / 2.0	308.6 / 0.7	261.2 / 0.5c
Refugees and others of concern to UNHCR (000)	630.6n	273.8n	402.1e
Infant mortality rateb,l (per 1 000 live births)	67.1	52.4	44.0c
Health: Total expenditure (% of GDP)	4.7	5.3	5.6d
Health: Physicians (per 1 000 pop.)	~0.0o	~0.0p	~0.0q
Education: Government expenditure (% of GDP)	4.6	4.6	3.5d
Education: Primary gross enrol. ratio (f/m per 100 pop.)	101.0 / 106.2	99.3 / 98.6	82.9 / 80.5c
Education: Secondary gross enrol. ratio (f/m per 100 pop.)	... / ...	27.5 / 34.8	30.8 / 33.7r
Education: Tertiary gross enrol. ratio (f/m per 100 pop.)	0.9 / 2.0h	1.9 / 2.4	2.5 / 4.9r
Intentional homicide rate (per 100 000 pop.)	...	8.5	7.0c
Seats held by women in national parliaments (%)	21.4	30.7	36.4

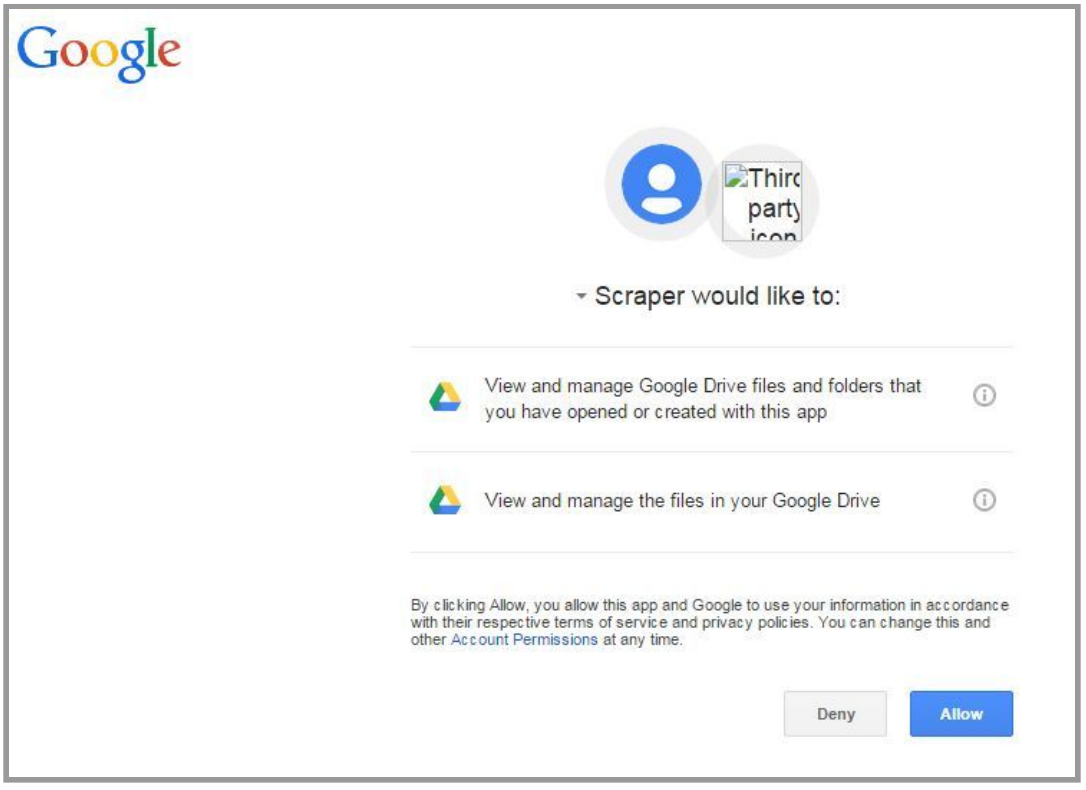
Export to Google Docs Option

You also have the option to export the HTML table to your Google Doc account. Here are the steps below.

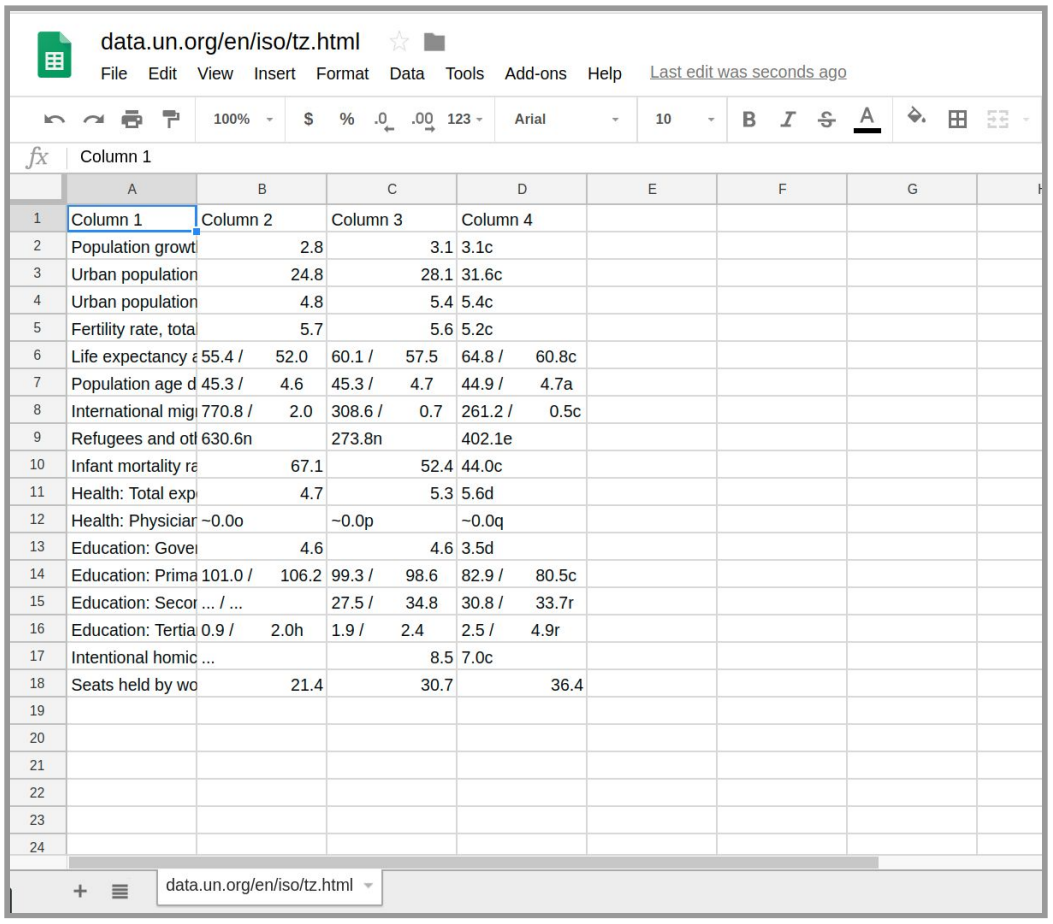
12. In the Scraper window that opens, click the **Export to Google Docs...** button.



13. 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 Scraper extension. Click **Allow**.



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



data.un.org/en/iso/tz.html

File Edit View Insert Format Data Tools Add-ons Help Last edit was seconds ago

Column 1

	A	B	C	D
1	Column 1	Column 2	Column 3	Column 4
2	Population growt	2.8	3.1	3.1c
3	Urban population	24.8	28.1	31.6c
4	Urban population	4.8	5.4	5.4c
5	Fertility rate, tota	5.7	5.6	5.2c
6	Life expectancy a	55.4 / 52.0	60.1 / 57.5	64.8 / 60.8c
7	Population age d	45.3 / 4.6	45.3 / 4.7	44.9 / 4.7a
8	International migi	770.8 / 2.0	308.6 / 0.7	261.2 / 0.5c
9	Refugees and otl	630.6n	273.8n	402.1e
10	Infant mortality ra	67.1	52.4	44.0c
11	Health: Total exp	4.7	5.3	5.6d
12	Health: Physiciar	~0.0o	~0.0p	~0.0q
13	Education: Gove	4.6	4.6	3.5d
14	Education: Prima	101.0 / 106.2	99.3 / 98.6	82.9 / 80.5c
15	Education: Secor	... / ...	27.5 / 34.8	30.8 / 33.7r
16	Education: Tertia	0.9 / 2.0h	1.9 / 2.4	2.5 / 4.9r
17	Intentional homic	...	8.5	7.0c
18	Seats held by wo	21.4	30.7	36.4

data.un.org/en/iso/tz.html

Exercise: Scrape Data Using Browser Extensions

Use the Chrome browser to scrape data from the "List of Regions" table available on this Wikipedia page: https://en.wikipedia.org/wiki/Regions_of_Tanzania

List of regions [edit]

Region	Capital	Districts	Area (km^2)	Population (2012)	Postcode	Zone	Map
Arusha Region	Arusha	7	37,576	1,694,310	23xxx	Northern	
Dar es Salaam Region	Dar es Salaam	5	1,393	4,364,541	11xxx	Coastal	
Dodoma Region	Dodoma	7	41,311	2,083,588	41xxx	Central	
Geita Region	Geita	5	20,054	1,739,530	30xxx	Lake	
Iringa Region	Iringa	5	35,503	941,238	51xxx	Southern Highlands	
Kagera Region	Bukoba	8	25,265	2,458,023	35xxx	Lake	
Katavi Region	Mpanda	3	45,843	564,604	50xxx	Western	
							

Task 3: Scraping Data Using ImportHTML

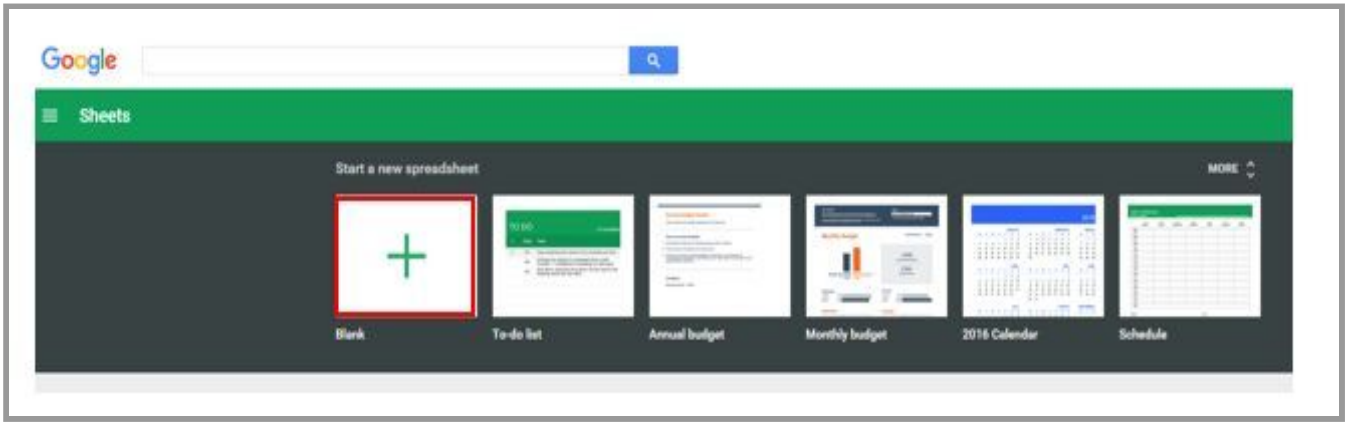
The Google Spreadsheet function **=importHTML(URL, QUERY, INDEX)** helps scrape a table from an HTML web page into a Google spreadsheet. Within this function:

- **URL**: should be the target web page containing the table and **MUST BE IN DOUBLE QUOTES**.
- **QUERY:** should have the string "table" and **MUST BE IN DOUBLE QUOTES**.
- **INDEX:** should be a number the identifies the N'th table in the page (counting starts at 0).

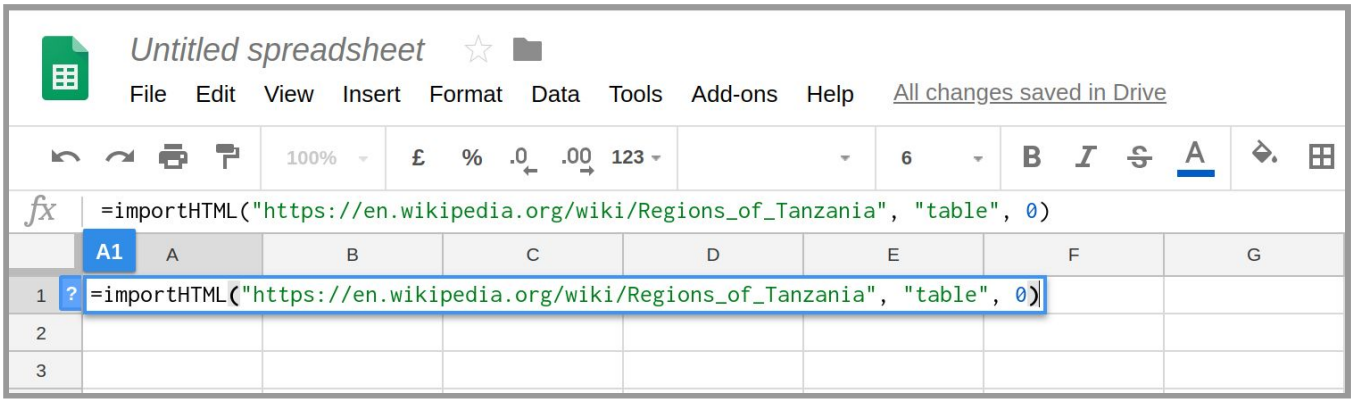
As an example, let's use Google Sheets to download data about the regions of Tanzania.

1. Open https://en.wikipedia.org/wiki/Regions_of_Tanzania and scroll down the page to find the **"List of regions"** table.

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



3. In the new Google Sheet that opens, type the following function: **=importHTML("https://en.wikipedia.org/wiki/Regions_of_Tanzania", "table", 0)**



4. Now press Enter and the Google Sheet populates data from the "List of regions' table from the Regions of Tanzania Wikipedia page.

Untitled spreadsheet

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

fx Arusha Region

	A	B	C	D	E	F	G	H	I
1	Region	Capital	Districts	Area (km2)	Population (2012	Postcode	Zone	Map	
2	Arusha Region	Arusha	7	37,576	1,694,310	23xxx	Northern		
3	Dar es Salaam R	Dar es Salaam	5	1,393	4,364,541	11xxx	Coastal		
4	Dodoma Region	Dodoma	7	41,311	2,083,588	41xxx	Central		
5	Geita Region	Geita	5	20,054	1,739,530	30xxx	Lake		
6	Iringa Region	Iringa	5	35,503	941,238	51xxx	Southern Highlands		
7	Kagera Region	Bukoba	8	25,265	2,458,023	35xxx	Lake		
8	Katavi Region	Mpanda	3	45,843	564,604	50xxx	Western		
9	Kigoma Region	Kigoma	8	37,040	2,127,930	47xxx	Western		
10	Kilimanjaro Regi	Moshi	7	13,250	1,640,087	25xxx	Northern		
11	Lindi Region	Lindi	6	66,040	864,652	65xxx	Coastal		
12	Manyara Region	Babati	6	44,522	1,425,131	27xxx	Northern		
13	Mara Region	Musoma	7	21,760	1,743,830	31xxx	Lake		
14	Mbeya Region	Mbeya	7	35,954	2,707,410[a]	53xxx	Southern Highlands		
15	Morogoro Regior	Morogoro	7	70,624	2,218,492	67xxx	Coastal		
16	Mtwara Region	Mtwara	7	16,710	1,270,854	63xxx	Coastal		
17	Mwanza Region	Mwanza	7	9,467	2,772,509	33xxx	Lake		
18	Njombe Region	Njombe	6	21,347	702,097	59xxx	Southern Highlands		
19	Pemba North Re	Wete	2	574	211,732	75xxx	Zanzibar		
20	Pemba South Re	Chake Chake	2	332	195,116	74xxx	Zanzibar		
21	Pwani Region	Kibaha	7	32,547	1,098,668	61xxx	Coastal		
22	Rukwa Region	Sumbawanga	4	22,792	1,004,539	55xxx	Southern Highlands		
23	Ruvuma Region	Songea	6	63,669	1,376,891	57xxx	Southern Highlands		
24	Shinyanga Regio	Shinyanga	5	18,901	1,534,808	37xxx	Lake		