

Web Scraping Tables using Pandas

Estimated Effort: 5 mins

The Pandas library in Python contains a function `read_html()` that can be used to extract tabular information from any web page.

Consider the following example:

Let us assume we want to extract the list of the largest banks in the world by market capitalization, from the following link:

```
1 URL = 'https://en.wikipedia.org/wiki/List_of_largest_banks'
```

We may use `pandas.read_html()` function in python to extract all the tables in the web page directly.

A snapshot of the webpage is shown below.

The screenshot shows the Wikipedia page titled "List of largest banks". The page includes a sidebar with navigation links like "Contents", "By market capitalization", "By total assets", and "Banks by country or territory". The main content area has a heading "List of largest banks" and a subheading "By market capitalization". Below this, there is a table listing the top 10 banks by market capitalization as of August 2023. The table has three columns: Rank, Bank name, and Market cap (US\$ billion). The banks listed are JPMorgan Chase, Bank of America, Industrial and Commercial Bank of China, Agricultural Bank of China, HDFC Bank, Wells Fargo, HSBC Holdings PLC, Morgan Stanley, China Construction Bank, and Bank of China.

Rank ↕	Bank name	Market cap ^[hide] (US\$ billion) ↕
1	JPMorgan Chase	419.25
2	Bank of America	231.52
3	Industrial and Commercial Bank of China	194.56
4	Agricultural Bank of China	160.68
5	HDFC Bank	157.91
6	Wells Fargo	155.87
7	HSBC Holdings PLC	148.90
8	Morgan Stanley	140.83
9	China Construction Bank	139.82
10	Bank of China	136.81

We can see that the required table is the first one in the web page.

Note: This is a live web page and it may get updated over time. The image shown above has been captured in November 2023. The process of data extraction remains the same.

We may execute the following lines of code to extract the required table from the web page.

```
1 import pandas as pd
2 URL = 'https://en.wikipedia.org/wiki/List_of_largest_banks'
3 tables = pd.read_html(URL)
4 df = tables[0]
5 print(df)
```

This will extract the required table as a dataframe `df`. The output of the print statement would look as shown below.

	Rank	Bank name	Market cap(US\$ billion)
0	1	JPMorgan Chase	419.25
1	2	Bank of America	231.52
2	3	Industrial and Commercial Bank of China	194.56
3	4	Agricultural Bank of China	160.68
4	5	HDFC Bank	157.91
5	6	Wells Fargo	155.87
6	7	HSBC Holdings PLC	148.90
7	8	Morgan Stanley	140.83
8	9	China Construction Bank	139.82
9	10	Bank of China	136.81

Although convenient, this method comes with its own set of limitations.

Firstly, web pages may have content saved in them as tables but they may not appear as tables on the web page.

For instance, consider the following URL showing the list of countries by GDP (nominal).

```
1 URL = 'https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)'
```

The images on the web page are also saved in tabular format. A snapshot of the web page is shared below.

List of countries by GDP (nominal)

75 languages

ArticleTalkReadView sourceView historyTools

From Wikipedia, the free encyclopedia

For countries by GDP based on purchasing power parity, see [List of countries by GDP \(PPP\)](#).

For countries by GDP per capita, see [List of countries by GDP \(nominal\) per capita](#).

Gross domestic product (GDP) is the market value of all final goods and services from a nation in a given year.^[1] Countries are sorted by nominal GDP estimates from financial and statistical institutions, which are calculated at market or government official exchange rates. Nominal GDP does not take into account differences in the cost of living in different countries, and the results can vary greatly from one year to another based on fluctuations in the exchange rates of the country's currency.^[1] Such fluctuations may change a country's ranking from one year to the next, even though they often make little or no difference in the standard of living of its population.^[6]

Comparisons of national wealth are also frequently made on the basis of purchasing power parity (PPP), to adjust for differences in the cost of living in different countries. Other metrics, nominal GDP per capita and a corresponding GDP (PPP) per capita are used for comparing national standard of living. On the whole, PPP per capita figures are less spread than nominal GDP per capita figures.^[6]

The rankings of national economies over time have changed considerably; the United States surpassed the British Empire's output around 1916,^[8] which in turn had surpassed the Qing dynasty in aggregate output decades earlier.^{[10][9]} Since China's transition to a socialist market economy through controlled privatisation and deregulation,^{[9][10]} the country has seen its ranking increase from ninth in 1978, to second in 2010, China's economic growth accelerated during this period and its share of global nominal GDP surged from 2% in 1980 to 18% in 2021.^{[11][12]} Among others, India has also experienced an economic boom since the implementation of economic liberalisation in the early 1990s.^[13]

The first list includes estimates compiled by the [International Monetary Fund's](#) [World Economic Outlook](#), the second list shows the [World Bank's](#) data, and the third list includes data compiled by the [United Nations Statistics Division](#). The IMF definitive data for the past year and estimates for the current year are published twice a year in April and October. Non-sovereign entities (the world, continents, and some *dependent territories*) and states with limited international recognition (such as [Kosovo](#) and [Taiwan](#)) are included in the list where they appear in the sources.

Country	GDP (trillion USD)
US	25.96
China	17.79
Germany	4.47
Japan	4.23
India	3.73
UK	3.33
France	3.16
Italy	2.19
Brazil	2.13
Canada	2.12

Color	Range (billion USD)
Dark Blue	> 20
Medium Blue	10-20
Light Blue	5-10
Very Light Blue	1-5
Dark Blue	750-1000
Medium Blue	500-750
Light Blue	250-500
Very Light Blue	100-250
Medium Blue	25-100
Light Blue	5-25
White	< 5

Table 1















Table 2

GDP (USD million) by country							
Country/Territory	UN region	IMF ^{[15][13]}		World Bank ^[14]		United Nations ^[11]	
		Forecast	Year	Estimate	Year	Estimate	Year
World	—	104,476,402	2023	100,662,011	2022	98,682,006	2021
United States	Americas	26,949,643	2023	25,462,700	2022	23,315,081	2021
China	Asia	17,700,899	^[11] 2023	17,963,171	^[11] 2022	17,734,131	^[11] 2021
Germany	Europe	4,429,838	2023	4,072,192	2022	4,259,935	2021
Japan	Asia	4,230,862	2023	4,231,141	2022	4,940,878	2021
India	Asia	3,732,224	2023	3,385,090	2022	3,201,471	2021
United Kingdom	Europe	3,332,059	2023	3,070,668	2022	3,131,378	2021
France	Europe	3,049,016	2023	2,782,905	2022	2,957,880	2021
Italy	Europe	2,186,082	2023	2,010,432	2022	2,107,703	2021
Brazil	Americas	2,126,805	2023	1,920,096	2022	1,608,981	2021
Canada	Americas	2,117,805	2023	2,129,840	2022	1,988,336	2021
Russia	Europe	1,862,470	2023	2,240,422	2022	1,778,782	2021
Mexico	Americas	1,811,468	2023	1,414,187	2022	1,272,839	2021
South Korea	Asia	1,709,232	2023	1,665,246	2022	1,810,968	2021
Australia	Oceania	1,687,713	2023	1,675,419	2022	1,734,532	2021
Spain	Europe	1,582,054	2023	1,397,509	2022	1,427,381	2021
Indonesia	Asia	1,417,387	2023	1,319,100	2022	1,186,093	2021
Turkey	Asia	1,154,800	2023	905,988	2022	819,034	2021
United States	Americas	1,000,315	2023	994,445	2022	6,549,847	2021

Table 3

Secondly, the contents of the tables in the web pages may contain elements such as hyperlink text and other denoters, which are also scraped directly using the pandas method. This may lead to a requirement of further cleaning of data. A closer look at table

3 in the image shown above indicates that there are many hyperlink texts which are also going to be treated as information by the pandas function.

GDP (USD million) by country								
	Country/Territory	UN region	IMF [1][13]	World Bank [14]	United Nations [15]			
			Forecast	Year	Estimate	Year	Estimate	Year
	World	—	104,476,432	2023	100,562,011	2022	96,698,005	2021
1	 United States	Americas	26,949,643	2023	25,462,700	2022	23,315,081	2021
2	 China	Asia	17,700,899	[n 1] 2023	17,963,171	[n 3] 2022	17,734,131	[n 1] 2021
3	 Germany	Europe	4,429,838	2023	4,072,192	2022	4,259,935	2021
4	 Japan	Asia	4,230,862	2023	4,231,141	2022	4,940,878	2021
5	 India	Asia	3,732,224	2023	3,385,090	2022	3,201,471	2021
6	 United Kingdom	Europe	3,332,059	2023	3,070,668	2022	3,131,378	2021
7	 France	Europe	3,049,016	2023	2,782,905	2022	2,957,880	2021
8	 Italy	Europe	2,186,082	2023	2,010,432	2022	2,107,703	2021
9	 Brazil	Americas	2,126,809	2023	1,920,096	2022	1,608,981	2021
10	 Canada	Americas	2,117,805	2023	2,139,840	2022	1,988,336	2021
11	 Russia	Europe	1,862,470	2023	2,240,422	2022	1,778,782	2021
12	 Mexico	Americas	1,811,468	2023	1,414,187	2022	1,272,839	2021
13	 South Korea	Asia	1,709,232	2023	1,665,246	2022	1,810,966	2021
14	 Australia	Oceania	1,687,713	2023	1,675,419	2022	1,734,532	2021
15	 Spain	Europe	1,582,054	2023	1,397,509	2022	1,427,381	2021

We can extract the table using the code shown below.

```

1 import pandas as pd
2 URL = 'https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)'
3 tables = pd.read_html(URL)
4 df = tables(2) # the required table will have index 2
5 print(df)

```

The output of the print statement is shown below.

	Country/Territory	UN region	IMF[1][13]		World Bank[14]		United Nations[15]	
	Country/Territory	UN region	Forecast	Year	Estimate	Year	Estimate	Year
0	World	-	104476432	2023	100562011	2022	96698005	2021
1	United States	Americas	26949643	2023	25462700	2022	23315081	2021
2	China	Asia	17700899	[n 1]2023	17963171	[n 3]2022	17734131	[n 1]2021
3	Germany	Europe	4429838	2023	4072192	2022	4259935	2021
4	Japan	Asia	4230862	2023	4231141	2022	4940878	2021
..
209	Palau	Oceania	267	2023	-	-	218	2021
210	Kiribati	Oceania	246	2023	223	2022	227	2021
211	Nauru	Oceania	150	2023	151	2022	155	2021
212	Montserrat	Americas	-	-	-	-	72	2021
213	Tuvalu	Oceania	63	2023	60	2022	60	2021

Note that the hyperlink texts have also been retained in the code output.

It is further prudent to point out, that this method exclusively operates only on tabular data extraction. BeautifulSoup library still remains the default method of extracting any kind of information from web pages.

Author(s)

[Abhishek Gagneja](#)