

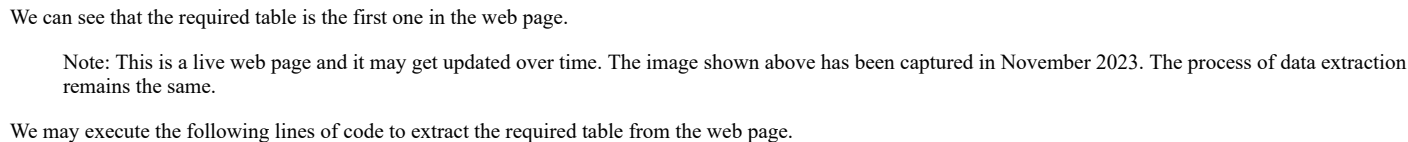
Estimated Effort: 5 mins

Consider the following example:

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

Copied!

A snapshot of the webpage is shown below.



1. 1
2. 2
3. 3

```
4. 4
5. 5

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)
```

Copied!

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. 1

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

Copied!

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

ArticleTalk

ReadView 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,^[9] which in turn had surpassed the Qing dynasty in aggregate output decades earlier.^[10] Since China's transition to a socialist market economy through controlled privatisation and deregulation,^[11] 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.^{[12][13]} Among others, India has also experienced an economic boom since the implementation of economic liberalisation in the early 1990s.^[14]

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.

28.96

17.79

4.43

4.03

3.73

3.59

3.06

2.79

2.13

2.10

ChinaUnited StatesJapanIndiaUKFranceGermanySouth KoreaTurkey

Largest economies in the world by GDP (nominal) in 2023 according to International Monetary Fund estimates^[15]

Countries by nominal GDP in 2019^[16]

< \$20 billion

\$20–50 billion

\$50–100 billion

\$100–200 billion

\$200–500 billion

\$500–1,000 billion

> \$1,000 billion

> \$5 trillion

Table

The table initially renais each country or territory with their latest available estimates, and can be renamed by either of the sources

The links in the "Country/Territory" row of the following table link to the article on the GDP or the economy of the respective country or territory.

| GDP (in \$B million) by country | | | | | | | | | |
|---------------------------------|-----------|---------------------|---------------------|----------------------------|---------------------|--------------------------------|---------------------|--|--|
| Country/Territory | UN region | IMF ^[17] | | World Bank ^[18] | | United Nations ^[19] | | | |
| | | Forecast | Year | Estimate | Year | Estimate | Year | | |
| World | --- | 104,478,432 | 2023 | 100,862,011 | 2022 | 98,882,006 | 2021 | | |
| 1 United States | Americas | 26,949,643 | 2023 | 25,462,700 | 2022 | 23,315,081 | 2021 | | |
| 2 China | Asia | 17,700,899 | ^[1] 2023 | 17,963,171 | ^[1] 2022 | 17,734,131 | ^[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 | 3,732,324 | 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 | 2,762,905 | 2022 | 2,957,880 | 2021 | | |
| 8 Italy | Europe | 2,186,062 | 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 | 1,778,762 | 2021 | | |
| 12 Mexico | Americas | 1,811,468 | 2023 | 1,414,187 | 2022 | 1,272,839 | 2021 | | |
| 13 South Korea | Asia | 1,709,232 | 2023 | 1,669,246 | 2022 | 1,810,968 | 2021 | | |
| 14 Australia | Oceania | 1,687,713 | 2023 | 1,675,419 | 2022 | 1,734,532 | 2021 | | |
| 15 Spain | Europe | 1,582,054 | 2023 | 1,397,009 | 2022 | 1,427,381 | 2021 | | |
| 16 Indonesia | Asia | 1,417,387 | 2023 | 1,319,100 | 2022 | 1,186,093 | 2021 | | |
| 17 Turkey | Asia | 1,154,600 | 2023 | 905,988 | 2022 | 819,034 | 2021 | | |
| 18 | | 1,053,945 | 2023 | 959,722 | 2022 | 1,043,045 | 2021 | | |

Table 1

Table 2

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.

about:blank

3/4

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. 1
2. 2
3. 3
4. 4
5. 5

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)
```

Copied!

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](#)