Google News Initiative

LESSON 15 : Part 01 of 04 in the Visualizing Data module

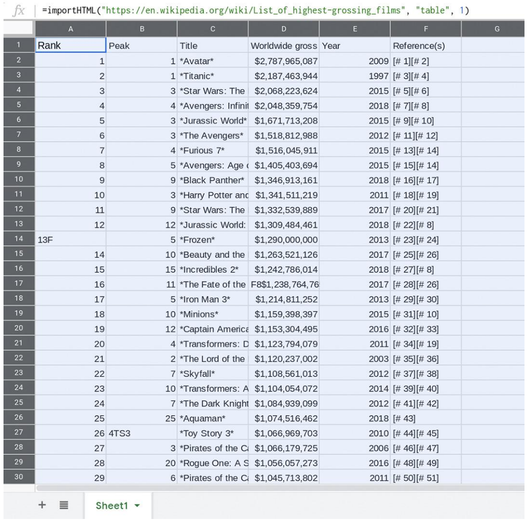
Google Sheets: Scraping data from the internet

Build your own data sets using Google Sheets.

Lesson overview

Learn to build your own data sets using Google Sheets.

There is a massive amount of data available on the internet that you can use to research and visualize stories. Finding the data, and getting it into a format you can work with is the first step.



```
=importHTML("https://en.wikipedia.org/wiki/List_of_highest-grossing_films", "table", 1)
```

	A	B	C	D	E	F	G
1	Rank	Peak	Title	Worldwide gross	Year	Reference(s)	
2	1	1	*Avatar*	$2,787,965,087	2009	[# 1][# 2]	
3	2	1	*Titanic*	$2,187,463,944	1997	[# 3][# 4]	
4	3	3	*Star Wars: The	$2,068,223,624	2015	[# 5][# 6]	
5	4	4	*Avengers: Infinit	$2,048,359,754	2018	[# 7][# 8]	
6	5	3	*Jurassic World*	$1,671,713,208	2015	[# 9][# 10]	
7	6	3	*The Avengers*	$1,518,812,988	2012	[# 11][# 12]	
8	7	4	*Furious 7*	$1,516,045,911	2015	[# 13][# 14]	
9	8	5	*Avengers: Age (	$1,405,403,694	2015	[# 15][# 14]	
10	9	9	*Black Panther*	$1,346,913,161	2018	[# 16][# 17]	
11	10	3	*Harry Potter anc	$1,341,511,219	2011	[# 18][# 19]	
12	11	9	*Star Wars: The	$1,332,539,889	2017	[# 20][# 21]	
13	12	12	*Jurassic World:	$1,309,484,461	2018	[# 22][# 8]	
14	13F	5	*Frozen*	$1,290,000,000	2013	[# 23][# 24]	
15	14	10	*Beauty and the	$1,263,521,126	2017	[# 25][# 26]	
16	15	15	*Incredibles 2*	$1,242,786,014	2018	[# 27][# 8]	
17	16	11	*The Fate of the	F8$1,238,764,76	2017	[# 28][# 26]	
18	17	5	*Iron Man 3*	$1,214,811,252	2013	[# 29][# 30]	
19	18	10	*Minions*	$1,159,398,397	2015	[# 31][# 10]	
20	19	12	*Captain America	$1,153,304,495	2016	[# 32][# 33]	
21	20	4	*Transformers: D	$1,123,794,079	2011	[# 34][# 19]	
22	21	2	*The Lord of the	$1,120,237,002	2003	[# 35][# 36]	
23	22	7	*Skyfall*	$1,108,561,013	2012	[# 37][# 38]	
24	23	10	*Transformers: A	$1,104,054,072	2014	[# 39][# 40]	
25	24	7	*The Dark Knight	$1,084,939,099	2012	[# 41][# 42]	
26	25	25	*Aquaman*	$1,074,516,462	2018	[# 43]	
27	26	4TS3	*Toy Story 3*	$1,066,969,703	2010	[# 44][# 45]	
28	27	3	*Pirates of the Ca	$1,066,179,725	2006	[# 46][# 47]	
29	28	20	*Rogue One: A S	$1,056,057,273	2016	[# 48][# 49]	
30	29	6	*Pirates of the Ca	$1,045,713,802	2011	[# 50][# 51]	

1 Starting a new spreadsheet.

2 Finding reliable data.

3 Importing data to Google Sheets.

4 Troubleshooting and error messages.

5 Displaying your data.

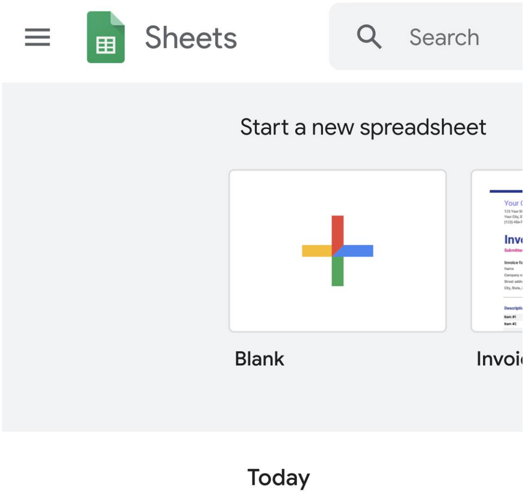
For more Data Journalism lessons, visit:

newsinitiative.withgoogle.com/training/course/data-journalism

Starting a new spreadsheet.

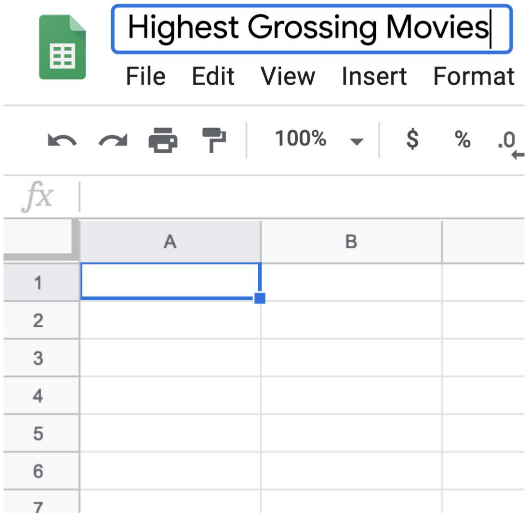
STEP 1 OF 2

First, you need to create a blank spreadsheet. Go to sheets.google.com. Under **Start a new spreadsheet**, click the **+** icon.



STEP 2 OF 2

To name your spreadsheet, click the text in the top left corner. Let's name this one "Highest Grossing Movies."

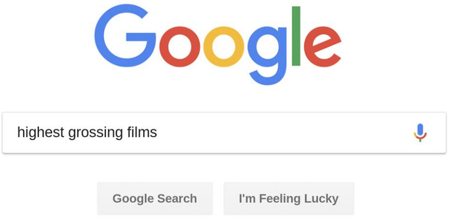


Finding reliable data.

By sourcing data from government sites, scientific publications, Wikipedia, Google Public Data Explorer and more, you can tell data stories on almost any topic. In this lesson, we'll practice with data about movies.

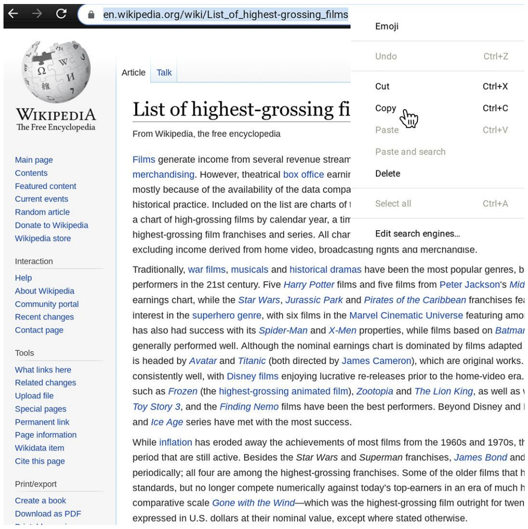
STEP 1 OF 2

Go to **google.com** and **search highest grossing films**. One of the first links should be a Wikipedia entry with multiple tables. One list, called "the top 50 highest-grossing films of all time" cites multiple references, so we will use that one. Always check to make sure you're scraping data from reliable sources.


highest grossing films

STEP 2 OF 2

To import this table to Google Sheets, copy the address of the Wikipedia page by highlighting the URL, right clicking on it, and selecting **copy**.


en.wikipedia.org/wiki/List_of_highest-grossing_films

List of highest-grossing fi

From Wikipedia, the free encyclopedia

Films generate income from several revenue stream
merchandising. However, theatrical box office earnir
mostly because of the availability of the data compa
historical practice. Included on the list are charts of t
a chart of high-grossing films by calendar year, a tim
highest-grossing film franchises and series. All char
excluding income derived from home video, broadcasting rights and merchandise.

Traditionally, war films, musicals and historical dramas have been the most popular genres, b
performers in the 21st century. Five *Harry Potter* films and five films from Peter Jackson's *Mid*
earnings chart, while the *Star Wars*, *Jurassic Park* and *Pirates of the Caribbean* franchises fea
interest in the superhero genre, with six films in the Marvel Cinematic Universe featuring amo
has also had success with its *Spider-Man* and *X-Men* properties, while films based on *Batma*
generally performed well. Although the nominal earnings chart is dominated by films adapted
is headed by *Avatar* and *Titanic* (both directed by James Cameron), which are original works.
consistently well, with Disney films enjoying lucrative re-releases prior to the home-video era.
such as *Frozen* (the highest-grossing animated film), *Zootopia* and *The Lion King*, as well as
Toy Story 3, and the *Finding Nemo* films have been the best performers. Beyond Disney and
and *Ice Age* series have met with the most success.

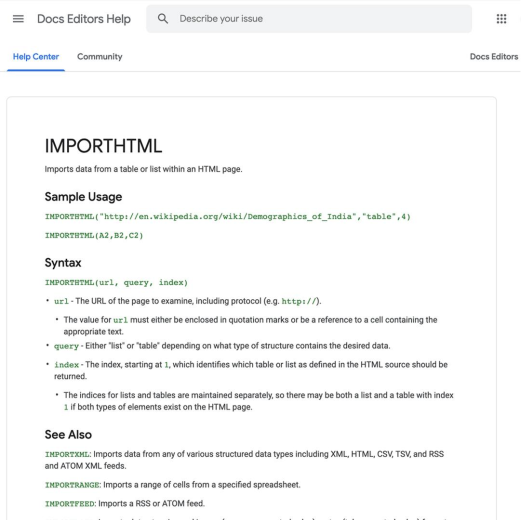
While inflation has eroded away the achievements of most films from the 1960s and 1970s, th
period that are still active. Besides the *Star Wars* and *Superman* franchises, *James Bond* and
periodically; all four are among the highest-grossing franchises. Some of the older films that h
standards, but no longer compete numerically against today's top-earners in an era of much h
comparative scale *Gone with the Wind*—which was the highest-grossing film outright for twen
expressed in U.S. dollars at their nominal value, except where stated otherwise.

Importing data to Google Sheets.

We'll use importHTML to import the table from Wikipedia to our spreadsheet. This powerful formula is built into Google Sheets to help you import tables or lists from web pages. To learn more about how importHTML works and see examples, read the Google Sheets documentation pages.

STEP 1 OF 3

The **`importHTML`** tool needs three parameters to work:
1) a URL
2) the type of data we're collecting, either a table or list
3) the number representing the position of the table or list in the HTML code.
In this example, the first instance of a table would be numbered as one, as the table we want is the first one that shows up in the HTML. You can use trial and error to find what the position of the table is (1, 2, 3, etc.) or right click the webpage, select **Inspect** > **Find** to locate the table in the code.



IMPORTHTML

Imports data from a table or list within an HTML page.

Sample Usage

```
IMPORTHTML("http://en.wikipedia.org/wiki/Demographics_of_India","table",4)
IMPORTHTML(A2,B2,C2)
```

Syntax

```
IMPORTHTML(url, query, index)
```

- `url` - The URL of the page to examine, including protocol (e.g. `http://`).
  - The value for `url` must either be enclosed in quotation marks or be a reference to a cell containing the appropriate text.
- `query` - Either "list" or "table" depending on what type of structure contains the desired data.
- `index` - The index, starting at `1`, which identifies which table or list as defined in the HTML source should be returned.
  - The indices for lists and tables are maintained separately, so there may be both a list and a table with index `1` if both types of elements exist on the HTML page.

See Also

`IMPORTXML`: Imports data from any of various structured data types including XML, HTML, CSV, TSV, and RSS and ATOM XML feeds.

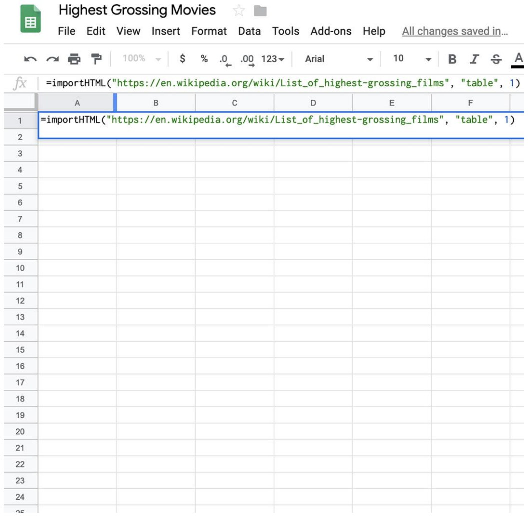
`IMPORTRANGE`: Imports a range of cells from a specified spreadsheet.

`IMPORTFEED`: Imports a RSS or ATOM feed.

STEP 2 OF 3

Go to the blank sheet you created and navigate to cell A1. Type:

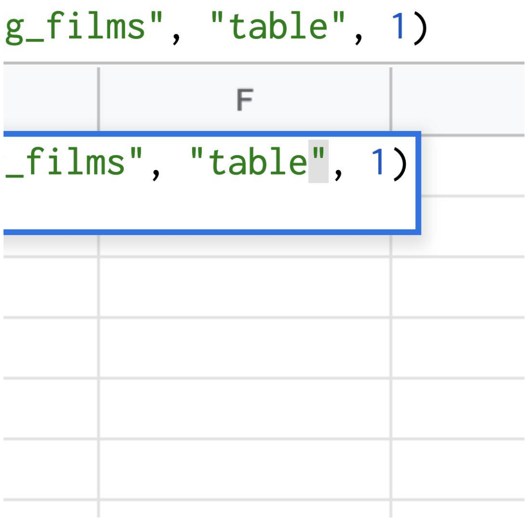
```
=importHTML("https://en.wikipedia.org/wiki/List_of_highest-grossing_films", "table", 1)
```



```
=importHTML("https://en.wikipedia.org/wiki/List_of_highest-grossing_films", "table", 1)
```

STEP 3 OF 3

Notice that the URL and the element type (in our case, table) go between quotes — this will make the parameters green. The last parameter is a number not within quotes and it will be colored blue.



```
g_films", "table", 1)
```

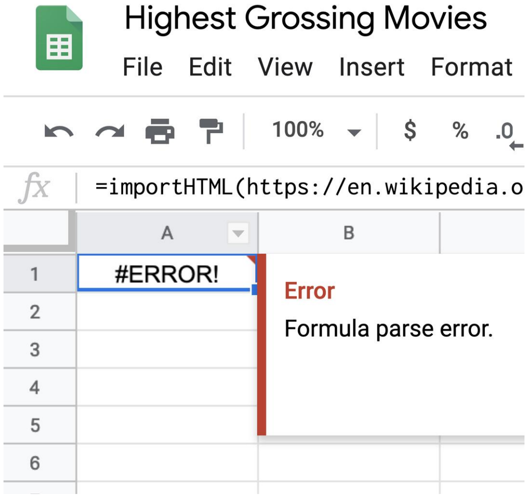
```
_films", "table", 1)
```

Troubleshooting and error messages.

SINGLE STEP

If you get an ERROR! Message, check to make sure the quotes are double quotes as shown in the example.

If you get a VALUE! error, check to make sure you don't have extra parentheses or quotation marks in the cell.



```
=importHTML(https://en.wikipedia.o
```

#ERROR!

Error

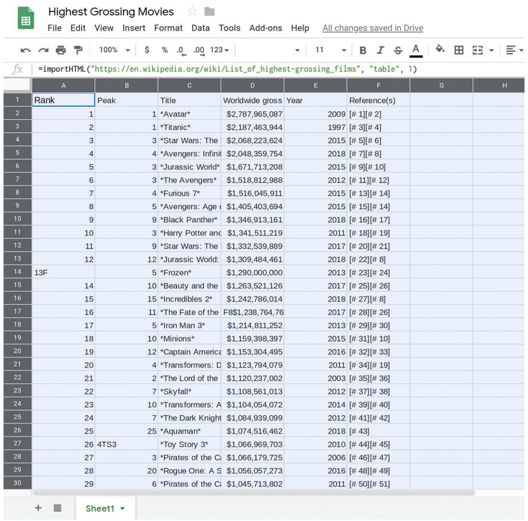
Formula parse error.

Displaying your data.

SINGLE STEP

Once your ImportHTML formula is correct, press **enter** and give Google Sheets a couple of seconds. The table should load with all the rows and columns formatted.

Notice that there are some elements we need to remove so that we can visualize this data. We will learn this in the next lesson, “Google Sheets: Cleaning data.”



Highest Grossing Movies
File Edit View Insert Format Data Tools Add-ons Help
All changes saved in Drive
100%
$ % .0 .00 123
11
B I S A
Sheet1

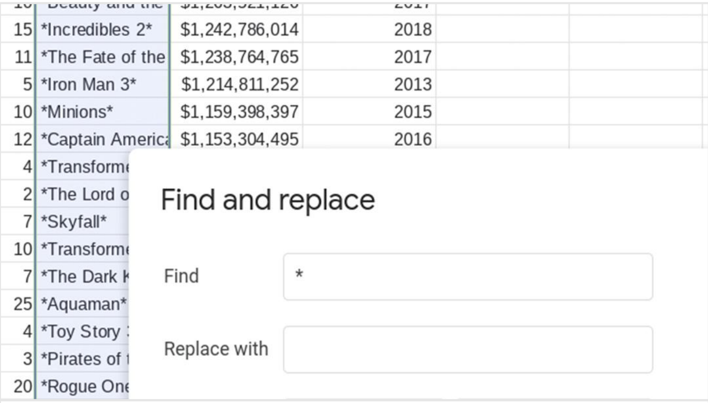
```
=importHTML("https://en.wikipedia.org/wiki/List_of_highest-grossing_films", "table", 1)
```

	A	B	C	D	E	F	G	H
1	Rank	Peak	Title	Worldwide gross	Year	Reference(s)		
2	1	1	*Avatar*	$2,787,965,087	2009	[# 1][# 2]		
3	2	1	*Titanic*	$2,187,463,944	1997	[# 3][# 4]		
4	3	3	*Star Wars: The	$2,068,223,624	2015	[# 5][# 6]		
5	4	4	*Avengers: Infinit	$2,048,359,754	2018	[# 7][# 8]		
6	5	3	*Jurassic World*	$1,671,713,208	2015	[# 9][# 10]		
7	6	3	*The Avengers*	$1,518,812,988	2012	[# 11][# 12]		
8	7	4	*Furious 7*	$1,516,045,911	2015	[# 13][# 14]		
9	8	5	*Avengers: Age (	$1,405,403,694	2015	[# 15][# 14]		
10	9	9	*Black Panther*	$1,346,913,161	2018	[# 16][# 17]		
11	10	3	*Harry Potter anc	$1,341,511,219	2011	[# 18][# 19]		
12	11	9	*Star Wars: The	$1,332,539,889	2017	[# 20][# 21]		
13	12	12	*Jurassic World:	$1,309,484,461	2018	[# 22][# 8]		
14	13F	5	*Frozen*	$1,290,000,000	2013	[# 23][# 24]		
15	14	10	*Beauty and the	$1,263,521,126	2017	[# 25][# 26]		
16	15	15	*Incredibles 2*	$1,242,786,014	2018	[# 27][# 8]		
17	16	11	*The Fate of the	F8$1,238,764,76	2017	[# 28][# 26]		
18	17	5	*Iron Man 3*	$1,214,811,252	2013	[# 29][# 30]		
19	18	10	*Minions*	$1,159,398,397	2015	[# 31][# 10]		
20	19	12	*Captain America	$1,153,304,495	2016	[# 32][# 33]		
21	20	4	*Transformers: D	$1,123,794,079	2011	[# 34][# 19]		
22	21	2	*The Lord of the	$1,120,237,002	2003	[# 35][# 36]		
23	22	7	*Skyfall*	$1,108,561,013	2012	[# 37][# 38]		
24	23	10	*Transformers: A	$1,104,054,072	2014	[# 39][# 40]		
25	24	7	*The Dark Knight	$1,084,939,099	2012	[# 41][# 42]		
26	25	25	*Aquaman*	$1,074,516,462	2018	[# 43]		
27	26	4TS3	*Toy Story 3*	$1,066,969,703	2010	[# 44][# 45]		
28	27	3	*Pirates of the C	$1,066,179,725	2006	[# 46][# 47]		
29	28	20	*Rogue One: A S	$1,056,057,273	2016	[# 48][# 49]		
30	29	6	*Pirates of the C	$1,045,713,802	2011	[# 50][# 51]		

Congratulations!

You completed “Google Sheets: Scraping data from the internet.”

To continue building your digital journalism skills and work toward Google News Initiative certification, go to our Training Center website and take another lesson:



10	Beauty and the	$1,263,521,126	2017
15	*Incredibles 2*	$1,242,786,014	2018
11	*The Fate of the	$1,238,764,765	2017
5	*Iron Man 3*	$1,214,811,252	2013
10	*Minions*	$1,159,398,397	2015
12	*Captain America	$1,153,304,495	2016
4	*Transforme		
2	*The Lord o		
7	*Skyfall*		
10	*Transforme		
7	*The Dark K		
25	*Aquaman*		
4	*Toy Story :		
3	*Pirates of t		
20	*Rogue One		

Find and replace

Find	*
Replace with	

Google Sheets: Cleaning data

Prepare your data for analysis and visualization.

For more Data Journalism lessons, visit:

newsinitiative.withgoogle.com/training/course/data-journalism