
LESSON 16 : Part 02 of 04 in the Visualizing Data module

Google Sheets: Cleaning data

Prepare your data for analysis and visualization.

Lesson overview

Learn to clean data in preparation for visualization.

In the previous lesson, [Google Sheets: Scraping data from the internet](#), we learned how to import a table from the Web using `importHTML`. In this lesson, we'll learn how to clean the data so it's ready for analysis and visualization.

5	*Frozen*	\$1,290,000,000	2013		
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	*Transform				
2	*The Lord o				
7	*Skyfall*				
10	*Transform				
7	*The Dark K				
25	*Aquaman*				
4	*Toy Story				
3	*Pirates of i				
20	*Rogue One				
6	*Pirates of i				
24	*Despicable				
1	*Jurassic P				
22	*Finding Do				
2	*Star Wars:				
5	*Alice in W				
24	*Zootopia*				
14	*The Hobbit				
4	*The Dark K				
2	*Harry Pott				

Find and replace

Find

Replace with

Search

Match case

Match entire cell contents

Search using regular expressions [Help](#)

Also search within formulas

- 1 Making data editable.
- 2 Editing the data.
- 3 Batch editing with **Find and replace**.

For more Data Journalism lessons, visit:

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

Making data editable.

“Cleaning data” means making it usable to work with: ensuring a table has integrity, is free from inconsistencies and is structured in a way that computers will understand. That means we will remove duplicate rows, delete undesired characters and ensure that columns hold only one type of data, for example numbers or text, but not both. First, we need to make the data editable.

STEP 1 OF 3

This table shows the result of importHTML. In this form, any changes to the data source (the Wikipedia page) will automatically be reflected here, and are updated at least once an hour. However, we can't edit the values in the cells to remove undesired characters. We will use **paste special** in Google Sheets to create a static snapshot of the data. With this, we will lose the ability to update the table automatically via importHTML, but we will be able to edit it.

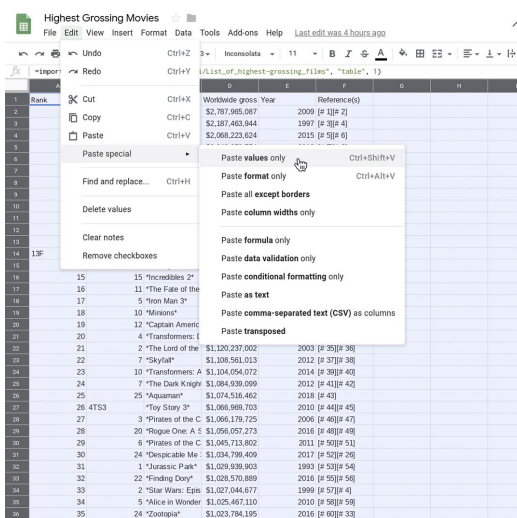
Rank	Peak	Title	Worldwide gross	Year	Reference(s)
1	1	*Avatar*	\$2,787,965,087	2009	[# 1][# 2]
2	2	*Titanic*	\$2,187,463,944	1997	[# 3][# 4]
3	3	*Star Wars: The	\$2,068,223,624	2015	[# 5][# 6]
4	4	*Avengers: Infi	\$2,048,359,754	2015	[# 7][# 8]
5	5	*Jurassic World*	\$1,671,713,208	2015	[# 9][# 10]
6	6	*The Avengers*	\$1,518,812,988	2012	[# 11][# 12]
7	7	*Furious 7*	\$1,516,045,911	2015	[# 13][# 14]
8	8	*Avengers: Age	\$1,495,403,894	2015	[# 15][# 16]
9	9	*Black Panther*	\$1,346,913,161	2018	[# 17][# 18]
10	10	*Harry Potter	\$1,341,511,219	2011	[# 19][# 20]
11	11	*Star Wars: The	\$1,332,539,899	2017	[# 21][# 22]
12	12	*Jurassic World	\$1,309,484,461	2018	[# 23][# 24]
13F	13	*Frozen*	\$1,290,000,000	2013	[# 25][# 26]
14	14	*Beauty and the	\$1,263,521,126	2017	[# 27][# 28]
15	15	*Incredibles 2*	\$1,242,786,014	2018	[# 29][# 30]
16	16	*The Fate of the	\$881,238,764,76	2017	[# 31][# 32]
17	17	*Iron Man 3*	\$1,214,811,252	2013	[# 33][# 34]
18	18	*Minions*	\$1,159,398,397	2015	[# 35][# 36]
19	19	*Captain Americ	\$1,153,304,495	2011	[# 37][# 38]
20	20	*Transformers: C	\$1,123,794,079	2003	[# 39][# 40]
21	21	*The Lord of the	\$1,120,237,002	2003	[# 41][# 42]
22	22	*Skyfall*	\$1,108,561,013	2012	[# 43][# 44]
23	23	*Transformers: A	\$1,104,054,072	2014	[# 45][# 46]
24	24	*The Dark Knigh	\$1,084,939,099	2012	[# 47][# 48]
25	25	*Aquaman*	\$1,074,516,462	2018	[# 49]
26	26	*Toy Story 3*	\$1,066,969,703	2010	[# 50][# 51]
27	27	*Pirates of the C	\$1,066,179,725	2006	[# 52][# 53]
28	28	*Rogue One: A S	\$1,056,057,273	2016	[# 54][# 55]
29	29	*Pirates of the C	\$1,045,713,802	2011	[# 56][# 57]

STEP 2 OF 3

Select all of the data by left-clicking in the top left rectangle in your sheet.

Once all cells are highlighted, click **Edit > Copy**.

Select **Edit > Paste special > Paste values only**. We're now able to edit the table.



STEP 3 OF 3

To make editing easier, we'll freeze the row with the names of the columns.

Hover the mouse cursor to the line just above row 1 over the gray bar. You will notice the cursor turns into a glove. Drag the bar to the bottom of row 1 and leave it there.

Now the top row is frozen.

The screenshot shows a spreadsheet titled "Highest Grossing Movies" with a menu bar (File, Edit, View, Insert, Format, Data, Tools, Add-ons, Help) and a toolbar. The spreadsheet has columns labeled A through H. The first row (row 1) is highlighted in blue and is frozen, as indicated by a small icon in the top-left corner of the cell. The data in the table is as follows:

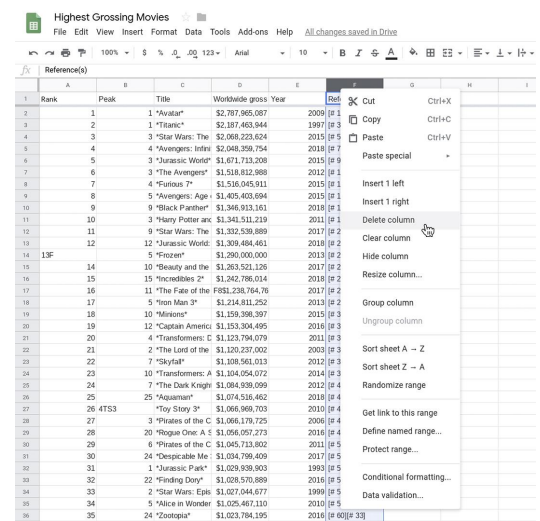
Rank	Peak	Title	Workwide gross	Year	Reference(s)
1	1	"Avatar"	\$2,787,965,087	2009 (# 1)(# 2)	
2	2	"Titanic"	\$2,187,463,944	1997 (# 3)(# 4)	
3	3	"Star Wars: The Force Awakens"	\$2,069,223,624	2015 (# 5)(# 6)	
4	4	"Avengers: Infinity War"	\$2,048,359,754	2018 (# 7)(# 8)	
5	5	"Jurassic World"	\$1,671,713,208	2015 (# 9)(# 10)	
6	6	"The Avengers"	\$1,518,812,888	2012 (# 11)(# 12)	
7	7	"Jurassic 2"	\$1,516,945,911	2015 (# 13)(# 14)	
8	8	"Avengers: Age of Ultron"	\$1,405,403,694	2015 (# 15)(# 16)	
9	9	"Black Panther"	\$1,346,913,161	2018 (# 16)(# 17)	
10	10	"Harry Potter and the Chamber of Secrets"	\$1,341,511,219	2011 (# 18)(# 19)	
11	11	"Star Wars: The Empire Strikes Back"	\$1,332,538,889	2017 (# 20)(# 21)	
12	12	"Jurassic World: Fallen Kingdom"	\$1,309,484,461	2018 (# 22)(# 24)	
13	13	"Frozen"	\$1,299,000,000	2013 (# 23)(# 24)	
14	14	"Beauty and the Beast"	\$1,263,521,126	2017 (# 25)(# 26)	
15	15	"Incredibles 2"	\$1,242,786,014	2018 (# 27)(# 28)	
16	16	"The Fate of the Furious"	\$651,238,744.76	2017 (# 28)(# 29)	
17	17	"Iron Man 3"	\$1,214,811,252	2013 (# 29)(# 30)	
18	18	"Minions"	\$1,159,398,397	2015 (# 31)(# 32)	
19	19	"Captain America: Civil War"	\$1,153,358,495	2016 (# 32)(# 33)	
20	20	"Transformers: C"	\$1,123,734,079	2011 (# 34)(# 35)	
21	21	"The Lord of the Rings: The Two Towers"	\$1,120,237,002	2003 (# 35)(# 36)	
22	22	"Skyline"	\$1,109,561,013	2012 (# 37)(# 38)	
23	23	"Transformers: A"	\$1,104,054,072	2014 (# 39)(# 40)	
24	24	"The Dark Knight"	\$1,084,939,099	2012 (# 41)(# 42)	
25	25	"Aquaman"	\$1,074,516,462	2018 (# 43)	
26	26	"Toy Story 3"	\$1,066,969,703	2010 (# 44)(# 45)	
27	27	"Pirates of the Caribbean: The Curse of the Black Pearl"	\$1,066,179,725	2006 (# 46)(# 47)	
28	28	"Rogue One: A Star Wars Story"	\$1,056,057,273	2016 (# 48)(# 49)	
29	29	"Pirates of the Caribbean: On Stranger Tides"	\$1,045,713,802	2011 (# 50)(# 51)	
30	30	"Despicable Me"	\$1,034,799,409	2017 (# 52)(# 53)	
31	31	"Jurassic Park"	\$1,029,939,903	1993 (# 53)(# 54)	
32	32	"Finding Dory"	\$1,028,570,889	2016 (# 55)(# 56)	
33	33	"Star Wars: Episode I - The Phantom Menace"	\$1,027,044,677	1999 (# 57)(# 4)	
34	34	"Alice in Wonderland"	\$1,025,461,110	2010 (# 58)(# 59)	
35	35	"The Exorcist"	\$1,019,794,194	2016 (# 59)(# 60)	

Editing the data.

importHTML will import leftover characters from the Wikipedia table that are useful for humans, but not computers. Let's remove them and make our table cleaner!

STEP 1 OF 3

Since we don't need column F for this exercise, right-click on the letter F at the top of the column and select **Delete**.



STEP 2 OF 3

There is a letter "F" next to number 13 in row A14, and a "TS3" next to number 4 in cell B27. We will remove these characters so that only the numbers 13 and 4 remain.

5	4	4	*Avengers: Infi	\$2,048,359,754	201
6	5	3	*Jurassic World*	\$1,671,713,208	201
7	6	3	*The Avengers*	\$1,518,812,988	201
8	7	4	*Furious 7*	\$1,516,045,911	201
9	8	5	*Avengers: Age	\$1,405,403,694	201
10	9	9	*Black Panther*	\$1,346,913,161	201
11	10	3	*Harry Potter anc	\$1,341,511,219	201
12	11	9	*Star Wars: The	\$1,332,539,889	201
13	12	12	*Jurassic World:	\$1,309,484,461	201
14	13F	5	*Frozen*	\$1,290,000,000	201
15	14	10	*Beauty and the	\$1,263,521,126	201
16	15	15	*Incredibles 2*	\$1,242,786,014	201
17	16	11	*The Fate of the	F8\$1,238,764,76	201
18	17	5	*Iron Man 3*	\$1,214,811,252	201
19	18	10	*Minions*	\$1,159,398,397	201
20	19	12	*Captain Americ	\$1,153,304,495	201
21	20	4	*Transformers: C	\$1,123,794,079	201
22	21	2	*The Lord of the	\$1,120,237,002	200
23	22	7	*Skyfall*	\$1,108,561,013	201
24	23	10	*Transformers: A	\$1,104,054,072	201
25	24	7	*The Dark Knight	\$1,084,939,099	201
26	25	25	*Aquaman*	\$1,074,516,462	201
27	26	4	*Toy Story 3*	\$1,066,969,703	201
28	27	3	*Pirates of the C	\$1,066,179,725	200
29	28	20	*Rogue One: A S	\$1,056,057,273	201
30	29	6	*Pirates of the C	\$1,045,733,802	201
31	30	24	*Despicable Me	\$1,034,799,409	201
32	31	1	*Jurassic Park*	\$1,029,939,903	199
33	32	22	*Finding Dory*	\$1,028,570,889	201
34	33	2	*Star Wars: Epi	\$1,027,044,677	199
35	34	5	*Alice in Wonder	\$1,025,487,110	201
36	35	24	*Zootopia*	\$1,023,784,195	201

STEP 3 OF 3

Remove the extra letters in cells B40 and B48, so that only 19 and 8 remain. Do the same in D17 to remove the leading "F8".

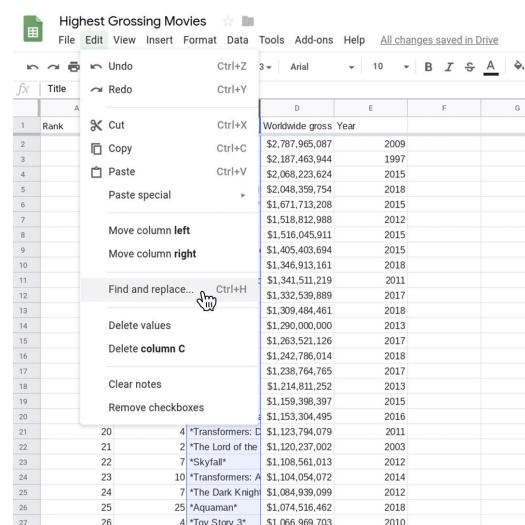
B	C	D	
19DM2	*Despicable Me :	\$970,761,885	
2	*The Lion King*	\$968,483,777	
30	*The Jungle Boo	\$966,550,600	
5	*Pirates of the Ca	\$963,420,425	
40	*Jumanji: Welcor	\$962,126,927	
10	*Harry Potter and	\$960,431,568	
24	*The Hobbit: The	\$958,366,855	
26	*The Hobbit: The	\$956,019,788	
8FN	*Finding Nemo*	\$940,335,536	
6	*Harry Potter and	\$940,018,451	
8	*Harry Potter and	\$934,546,568	

Batch editing with Find and replace.

Now, take a look at column C. Let's remove the leading and trailing * characters in a batch, rather than row by row, using the **Find and replace** feature.

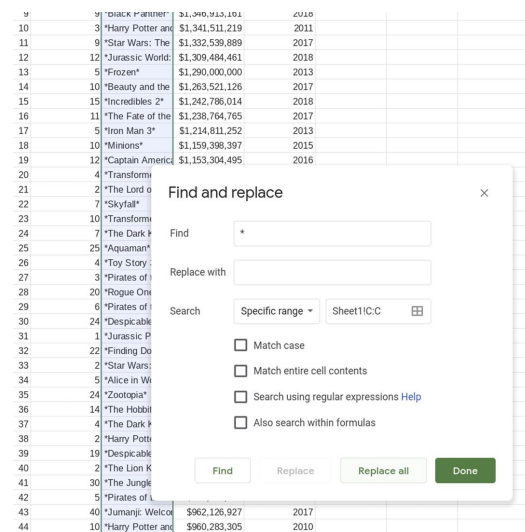
STEP 1 OF 5

Select column C by left-clicking on the letter C at the top of the column. Select **Edit > Find and replace**.



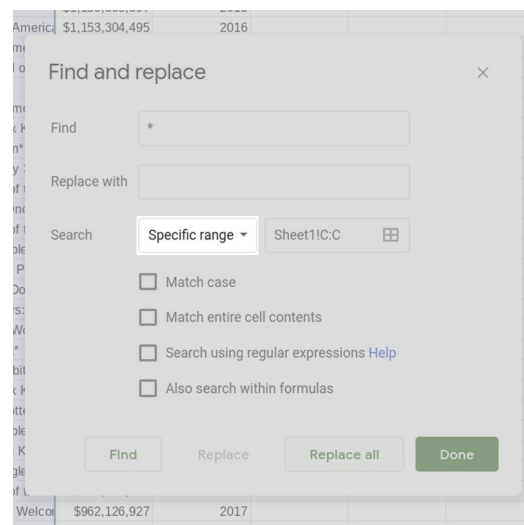
STEP 2 OF 5

In the first text box type the asterisk symbol: * (that's the character we want to find in column C). Leave the **Replace with** text box empty so that the asterisks get replaced with nothing, which means they will be deleted.



STEP 3 OF 5

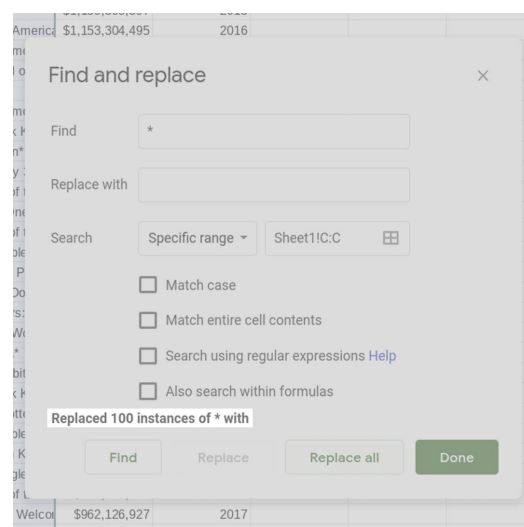
Make sure the **Search** option says **Specific range** and the range reflects the column you just selected. Leave the checkboxes unchecked.



STEP 4 OF 5

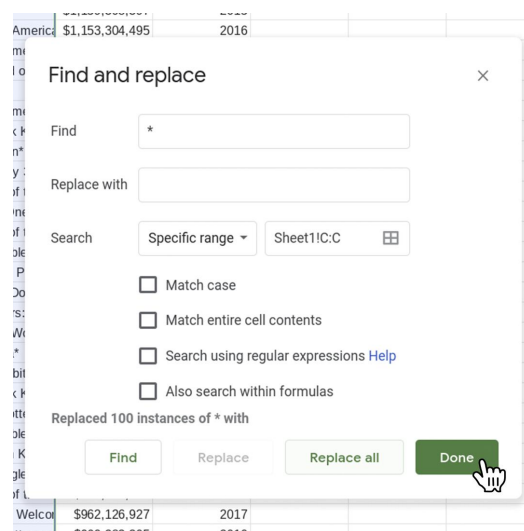
Select **Replace all**.

Notice Google Sheets will tell you it **Replaced 100 instances of * with** (nothing). That means you successfully removed 100 characters in 50 rows with just a few clicks!



STEP 5 OF 5

Select **Done**. Our table is now clean and ready for us to work with. In the next lesson, we will produce visualizations and get insights from the data.



Congratulations!

You completed “Google Sheets: Cleaning data.”

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

Rank	Movie Title	Worldwide Gross (\$)	Year
1	Avatar	\$2,043,600,000	2010
2	Avengers: Infinity War	\$678,812,000	2019
3	Furious 7	\$1,516,047,000	2015
4	Harry Potter and the Deathly Hallows - Part 2	\$1,511,361,000	2011
5	Pirates of the Caribbean: On Stranger Tides	\$963,420,425	2011
6	Jumanji: Welcome to the Jungle	\$962,126,927	2017
7	Harry Potter and the Chamber of Secrets	\$960,283,305	2002
8	The Fate of the Furious	\$1,120,237,002	2017
9	Skyfall	\$1,108,561,013	2012
10	Transformers: Age of Extinction	\$1,104,054,072	2014

Google Sheets: Visualizing data

Learn to build visualizations that help you interpret the data and tell data-driven stories.

For more Data Journalism lessons, visit:

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