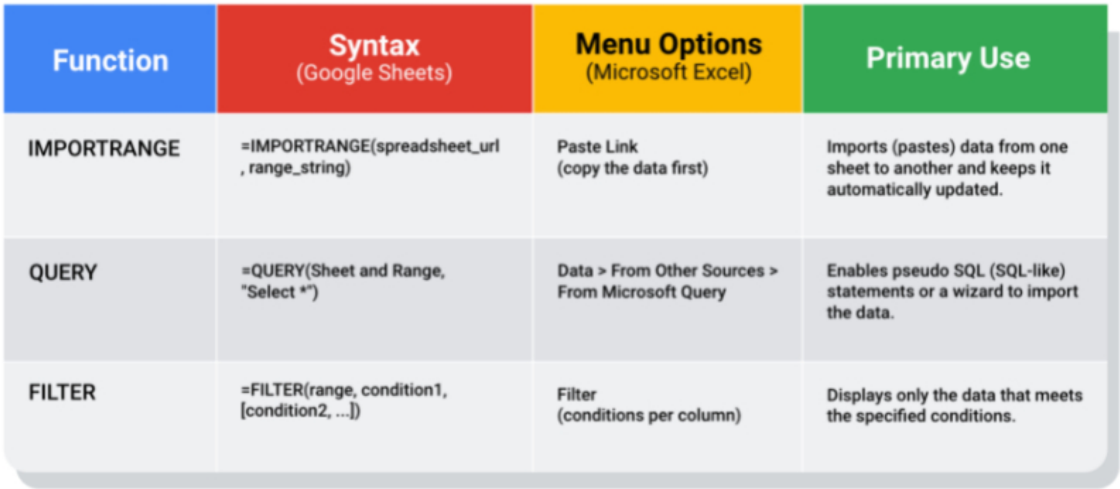
Advanced functions for speedy data cleaning

In this reading, you will learn about some advanced functions that can help you speed up the data cleaning process in spreadsheets. Below is a table summarizing three functions and what they do:

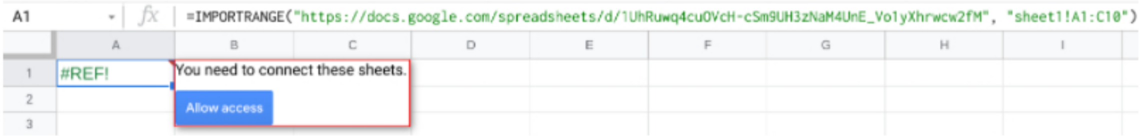


Function	Syntax (Google Sheets)	Menu Options (Microsoft Excel)	Primary Use
IMPORTRANGE	=IMPORTRANGE(spreadsheet_url, range_string)	Paste Link (copy the data first)	Imports (pastes) data from one sheet to another and keeps it automatically updated.
QUERY	=QUERY(Sheet and Range, "Select *")	Data > From Other Sources > From Microsoft Query	Enables pseudo SQL (SQL-like) statements or a wizard to import the data.
FILTER	=FILTER(range, condition1, [condition2, ...])	Filter (conditions per column)	Displays only the data that meets the specified conditions.

Keeping data clean and in sync with a source

The **IMPORTRANGE** function in Google Sheets and the **Paste Link** feature (a Paste Special option in Microsoft Excel) both allow you to insert data from one sheet to another. Using these on a large amount of data is more efficient than manual copying and pasting. They also reduce the chance of errors being introduced by copying and pasting the wrong data. They are also helpful for data cleaning because you can "cherry pick" the data you want to analyze and leave behind the data that isn't relevant to your project. Basically, it is like canceling noise from your data so you can focus on what is most important to solve your problem. This functionality is also useful for day-to-day data monitoring; with it, you can build a tracking spreadsheet to share the relevant data with others. The data is synced with the data source so when the data is updated in the source file, the tracked data is also refreshed.

If you are using IMPORTRANGE in Google sheets, data can be pulled from another spreadsheet, but you must allow access to the spreadsheet the first time it pulls the data. **The URL shown below is for syntax purposes only. Don't enter it in your own spreadsheet.** Replace it with a URL to a spreadsheet you have created so you can control access to it by clicking the Allow access button.



A1 =IMPORTRANGE("https://docs.google.com/spreadsheets/d/1UhRuwq4cuOVcH-cSm9UH3zNaM4UnE_VolyXhrwcw2fM", "sheet1!A1:C10")

#REF!

You need to connect these sheets.

Refer to the Google support page for IMPORTRANGE for the sample usage and syntax.

Example of using IMPORTRANGE

An analyst monitoring a fundraiser needs to track and ensure that matching funds are distributed. They use **IMPORTRANGE** to pull all the matching transactions into a spreadsheet containing all of the individual donations. This enables them to determine which donations eligible for matching funds still need to be processed. Because the total number of matching transactions increases daily, they simply need to change the range used by the function to import the most up-to-date data.

On Tuesday, they use the following to import the donor names and matched amounts:

=IMPORTRANGE(“https://docs.google.com/spreadsheets/d/1cOsHnBDzm9tBb8Hk_aLYfq3-o5FZ6DguPYRJ57992_Y”, “Matched Funds!A1:B4001”)

On Wednesday, another 500 transactions were processed. They increase the range used by 500 to easily include the latest transactions when importing the data to the individual donor spreadsheet:

=IMPORTRANGE(“https://docs.google.com/spreadsheets/d/1cOsHnBDzm9tBb8Hk_aLYfq3-o5FZ6DguPYRJ57992_Y”, “Matched Funds!A1:B4501”)

Note: The above examples are for illustrative purposes only. Don't copy and paste them into your spreadsheet. To try it out yourself, you will need to substitute your own URL (and sheet name if you have multiple tabs) along with the range of cells in the spreadsheet that you have populated with data.

Pulling data from other data sources

The **QUERY** function is also useful when you want to pull data from another spreadsheet. The **QUERY** function's SQL-like ability can extract specific data within a spreadsheet. For a large amount of data, using the QUERY function is faster than filtering data manually. This is especially true when repeated filtering is required. For example, you could generate a list of all customers who bought your company’s products in a particular month using manual filtering. But if you also want to figure out customer growth month over month, you have to copy the filtered data to a new spreadsheet, filter the data for sales during the following month, and then copy those results for the analysis. With the **QUERY** function, you can get all the data for both months without a need to change your original dataset or copy results.

The **QUERY** function syntax is similar to **IMPORTRANGE**. You enter the sheet by name and the range of data that you want to query from, and then use the SQL **SELECT** command to select the specific columns. You can also add specific criteria after the **SELECT** statement by including a **WHERE** statement. But remember, all of the SQL code you use has to be placed between the quotes!

Google Sheets run the Google Visualization API Query Language across the data. Excel spreadsheets use a query wizard to guide you through the steps to connect to a data source and select the tables. In either case, you are able to be sure that the data imported is verified and clean based on the criteria in the query.

Examples of using QUERY

Check out the Google support page for the QUERY function with sample usage, syntax, and examples you can download in a Google sheet.

Link to make a copy of the sheet: QUERY examples

Real life solution

Analysts can use SQL to pull a specific dataset into a spreadsheet. They can then use the **QUERY** function to create multiple tabs (views) of that dataset. For example, one tab could contain all the sales data for a particular month and another tab could contain all the sales data from a specific region. This solution illustrates how SQL and spreadsheets are used well together.

Filtering data to get what you want

The **FILTER** function is fully internal to a spreadsheet and doesn’t require the use of a query language. The FILTER function lets you view only the rows (or columns) in the source data that meet your specified conditions. It makes it possible to pre-filter data before you analyze it.

The **FILTER** function might run faster than the **QUERY** function. But keep in mind, the **QUERY** function can be combined with other functions for more complex calculations. For example, the **QUERY** function can be used with other functions like **SUM** and **COUNT** to summarize data, but the **FILTER** function can't.

Example of using FILTER

Check out the Google support page for the FILTER function with sample usage, syntax, and examples you can download in a Google sheet.

Link to make a copy of the sheet: FILTER examples

✓ Completed   Go to next item

Like   Dislike   Report an issue