

The Excelguru Blog

More geeky stuff from the author of www.excelguru.ca...

Identify Duplicates Using Power Query

Posted on [December 9, 2015](#) by [Ken Puls](#)

Some time ago I got an email from Alex asking me if there was a way to identify duplicates using Power Query, but without removing non-duplicate records in the process. This post explores how to do that.

Background

Suppose someone has given you a list like the one shown below (which you can [download here](#) if you'd like to follow along):

	A	B
1	SKU Number	Brand
2	510007	Budweiser
3	510010	Canadian
4	510014	Canterbury
5	510037	Guinness
6	510039	Heineken
7	511046	Kokanee
8	510057	Miller
9	510059	OK Springs
10	510065	OK Springs
11	512032	Granville Island
12	510098	Guinness
13	510010	Canadian
14	510019	Corona Extra
15	510021	Corona Grande
16	510032	Granville Island
17	510033	Guinness
18	510038	Heineken
19	510032	Granville Island

While multiple brands are okay here, we need a list that shows only unique SKU numbers. While the list provided to you was supposed to be duplicate free, you're not 100% sure that it actually is. While it would be easy to just hit the SKU column with the Remove Duplicates function, you don't want to do that. Instead you'd like to identify which records have duplicate entries in the list.

So how do we do this?

Naturally, there will be a few different ways to do this. I'm carving off one method that is the easiest to replicate via the user interface...

Step 1: Link to the Data

Of course we'll start by pulling the data in to Power Query

- Click anywhere in the Products Table
- Create a new query → From Table

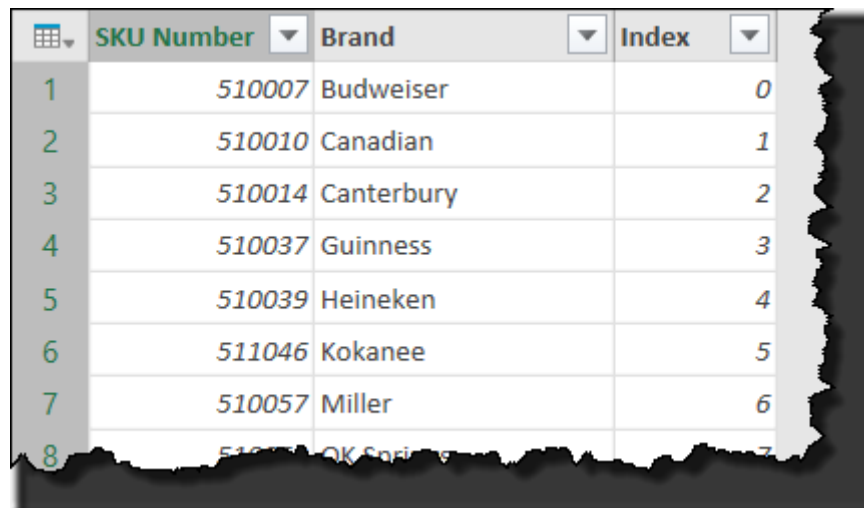
The data will be loaded in to Power Query, and you'll see two steps in the Applied Steps window:

- Source (pointing to your source data)
- Changed Type (setting the data types for the columns)

This might seem like an odd step right now, but we're going to add a Index column to this table as well. The reason will become apparent later, but for now:

- To to Add Column → Add Index Column → From 0

Your data should now look like this:



	SKU Number	Brand	Index
1	510007	Budweiser	0
2	510010	Canadian	1
3	510014	Canterbury	2
4	510037	Guinness	3
5	510039	Heineken	4
6	511046	Kokanee	5
7	510057	Miller	6
8	510057	OK Spring	7

Now we need to figure out how to flag any repeating SKU as a duplicate.

Step 2: Identify Duplicates via Grouping Rows

The trick here is to use the Group By feature in Power Query, while preserving the relevant matching records.

NOTE: We cover the Grouping feature in Chapter 14 of [M is for Data Monkey](#).

Here's how we do this:

- Go to Transform → Group By
- Set your Group By Options as follows:
 - Group By: SKU Number
 - New column name: Duplicates → Count Rows

Next, click the + to the right of the “New Column Name” section to add another detail row. Set it up as follows:

- New column name: Duplicates → All Rows

When you're done, the dialog should look like this:

×

Group By

Specify the columns to group by.

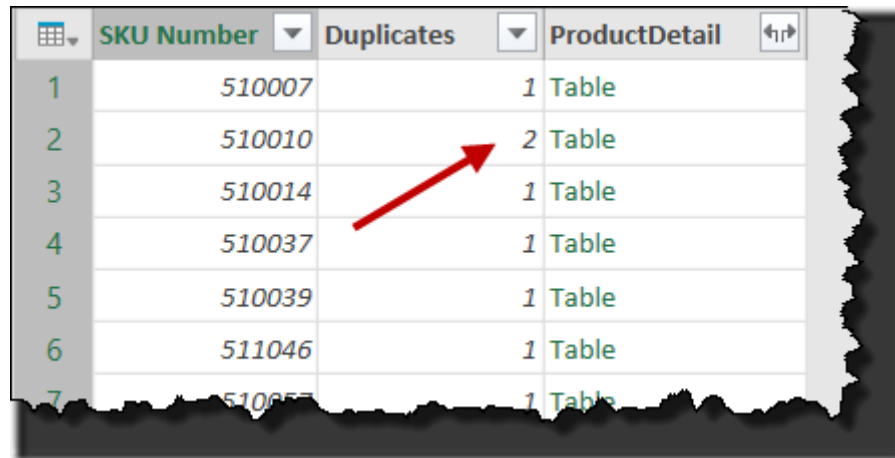
Group by +

SKU Number -

New column name	Operation	Column	+
Duplicates	Count Rows ▼	▼	-
ProductDetail	All Rows ▼	▼	-

OK Cancel

And upon clicking OK, the results will show that there are, indeed, items that show up more than once:

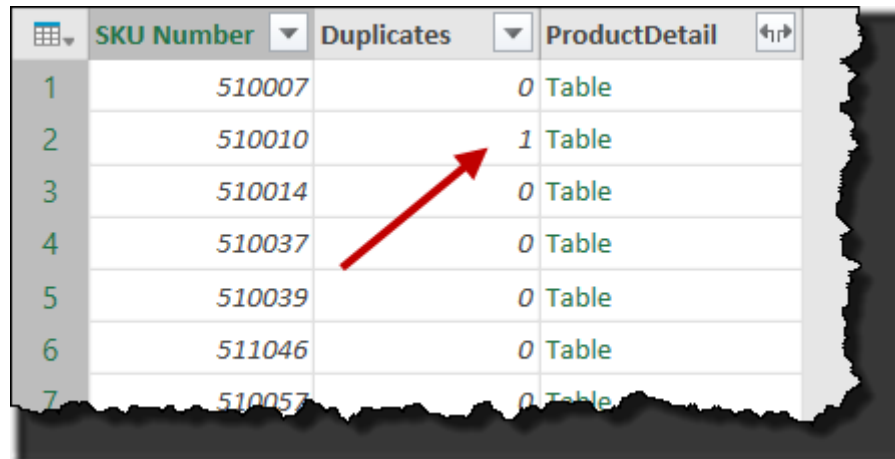


	SKU Number	Duplicates	ProductDetail
1	510007	1	Table
2	510010	2	Table
3	510014	1	Table
4	510037	1	Table
5	510039	1	Table
6	511046	1	Table
7	510057	1	Table

Let's tweak this a bit, and subtract 1 from each value. That would give us a truer representation as to how many duplicates there are.

- Select the Duplicates column -> Transform -> Subtract -> 1

Resulting in the following:



	SKU Number	Duplicates	ProductDetail
1	510007	0	Table
2	510010	1	Table
3	510014	0	Table
4	510037	0	Table
5	510039	0	Table
6	511046	0	Table
7	510057	0	Table

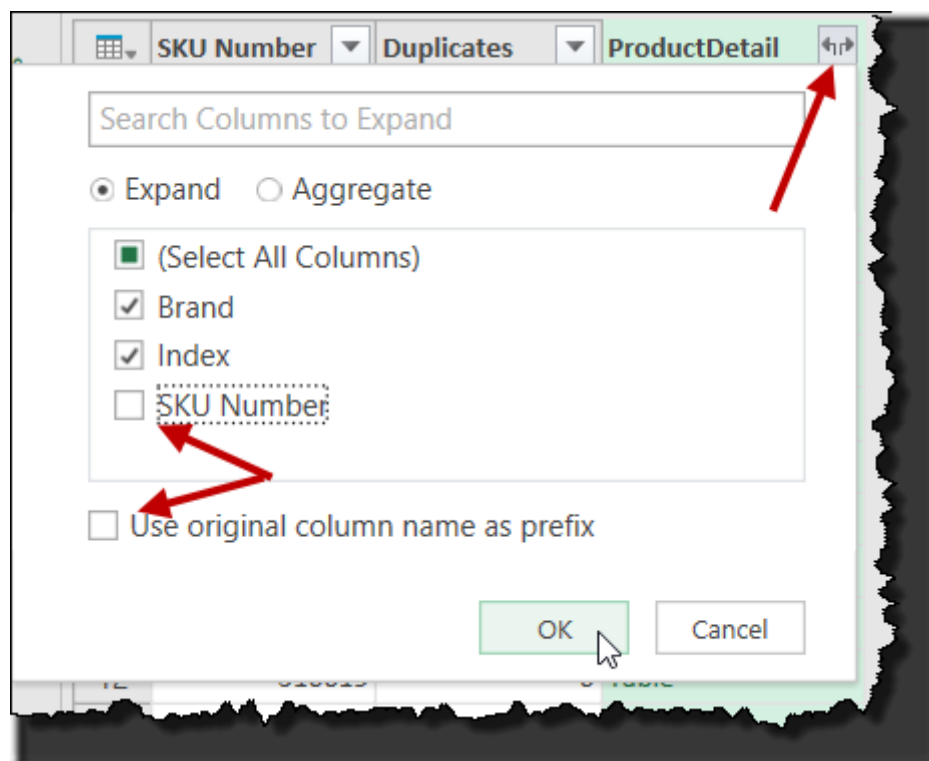
Much better. We're now seeing that SKU 510010 appears to have 1 duplicate entry in the data set.

But there is still an issue here. When we grouped our records, we lost both the Brand names column, but also any duplicate records. Since the whole point of this exercise was to Identify Duplicates but not remove the duplicate records, we're still not in a good place.

Step 3: Identify Duplicates and Show Duplicate Records

Let's fix this. Remember how we added a new step to show "All Rows" for the ProductDetail column? That step gave us the ability to do something pretty cool... it gave us the ability to get back all the lost records and product detail information we're currently missing.

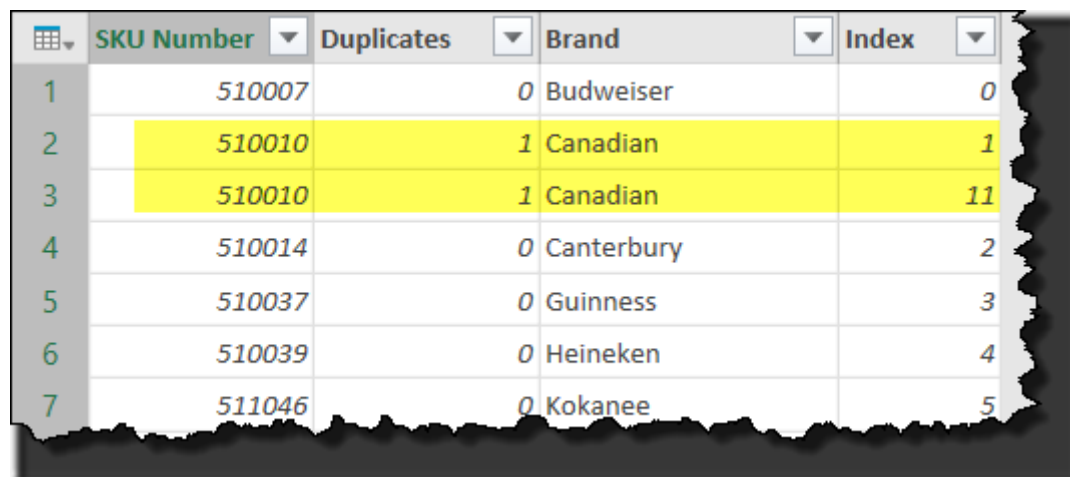
- Click the Expand button at the top right of the ProductDetail column
- Uncheck the SKU Number option (as we already have it)
- Uncheck the option to "Use original column name as prefix"



As you can see, this will bring back all the details we lost earlier.

Step 4: Final Cleanup

But hang on a second. Let's look at this output a bit more closely...



	SKU Number	Duplicates	Brand	Index
1	510007	0	Budweiser	0
2	510010	1	Canadian	1
3	510010	1	Canadian	11
4	510014	0	Canterbury	2
5	510037	0	Guinness	3
6	510039	0	Heineken	4
7	511046	0	Kokanee	5

Notice, that it re-sorted the data. That's not exactly a desirable outcome, as we are trying to flag duplicates for a reason. Maybe we want to know where they exist in an inventory count or we have some other reason for wanting to preserve the original sort order of our data. It's for this reason that we added the Index column earlier. That came through with the All Rows step, so let's put our data back into its original order.

- Click the drop down arrow on the Index column → Sort Ascending
- Right click the Index column → Remove

And we can now finalize the query:

- Rename the query to ShowDuplicates
- Go to Home → Close & Load

Step 5: Make the Duplicates Obvious

With the data now in an Excel table, we can make the duplicates even more obvious by applying some conditional formatting to the table. To do this:

- Select all the values in the Duplicates column of the table

- Go to Home → Conditional Formatting → Data Bars → Choose a colour

I chose blue data bars, which makes the data look like this:

SKU Number	Duplicates	Brand
510007	0	Budweiser
510010	1	Canadian
510014	0	Canterbury
510037	0	Guinness
510039	0	Heineken
511046	0	Kokanee
510057	0	Miller
510059	0	OK Springs
510065	0	OK Springs
512032	0	Granville Island
510098	0	Guinness
510010	1	Canadian
510019	0	Corona Extra
510021	0	Corona Grande
510032	1	Granville Island
510033	0	Guinness
510038	0	Heineken
510032	1	Granville Island

Conclusion

Our goal is now complete. We were able to identify duplicates and flag them without removing non-duplicate items. In addition, we have preserved the original order of the data in case that was important to us for any reason.

This entry was posted in [Excel](#), [Excel Add-ins](#), [General](#), [Get & Transform](#), [Office 2010](#), [Office 2013](#), [Office 2016](#), [Power Query](#) by [Ken Puls](#). Bookmark the [permalink \[https://www.excelguru.ca/blog/2015/12/09/identify-duplicates-using-power-query/\]](https://www.excelguru.ca/blog/2015/12/09/identify-duplicates-using-power-query/).