Data profiling in Excel Power Query

by George Mount

Power Query is rightly known for its abilities to *clean* data in Excel. But did you know about its data *profiling* powers? Let's take a look now at a modified version of the *penguins* dataset.

Download the demo file here.

Learn how to import a table into Excel Power Query here.

WHAT IS DATA PROFILING?

What is data profiling?

Think of it as your first look at the data from 30,000 feet. It allows you to understand its basic content, relationships and issues. Some questions you might ask as part of profiling:

- How accurate is this data? Is there anything obviously wrong with it?
- Is it clear what every variable and observation is supposed to be measuring?
- Do we have all the data we think we'll need to answer this question? Is there data missing?
- If the data's in Excel, are there formula errors that could impact our work?
- Has all of the data been transcribed correctly?

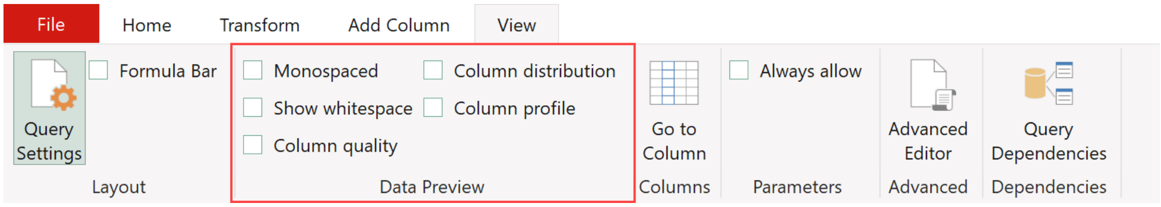
WHAT IS DATA PROFILING?

Data profiling in Power Query

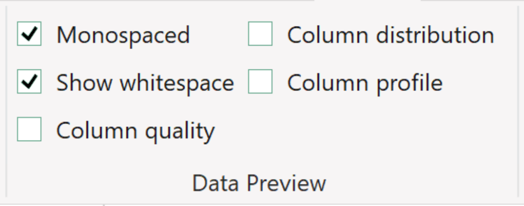
It makes sense that Power Query would include some data profiling features. After all, how can you clean the data if you don't understand what makes it dirty?

That said, data profiling is somewhat off the beaten path in the Power Query Editor: head to the **View** tab of the ribbon to find it in the "Data Preview" group:

Let's walk through what toggling on each of these five options will do.



SPACING OPTIONS



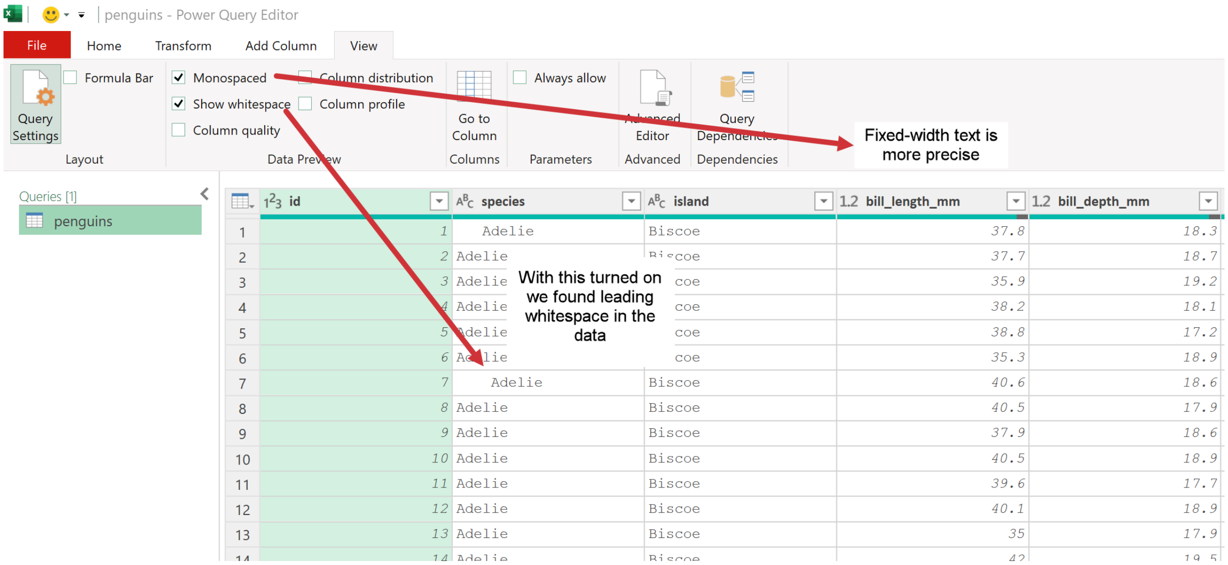
☑ Monospaced □ Column distribution
☑ Show whitespace □ Column profile
□ Column quality

Data Preview

"Monospaced" and "Show whitespace"

Start by checking on the first two options. These will change the appearance of the data in the Power Query editor:

- "Monospaced" will render the data as fixed-width text.
- "Show whitespace will render any leading spaces in the data.



☑ Monospaced
☑ Show whitespace

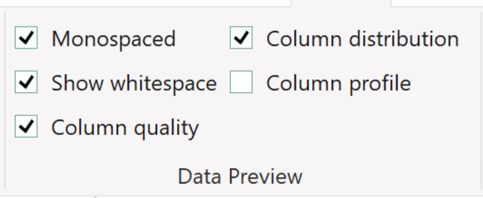
Fixed-width text is more precise

With this turned on we found leading whitespace in the data

	id	species	island	bill_length_mm	bill_depth_mm
1	1	Adelie	Biscoe	37.8	18.3
2	2	Adelie	Biscoe	37.7	18.7
3	3	Adelie	coe	35.9	19.2
4	4	Adelie	coe	38.2	18.1
5	5	Adelie	coe	38.8	17.2
6	6	Adelie	coe	35.3	18.9
7	7	Adelie	Biscoe	40.6	18.6
8	8	Adelie	Biscoe	40.5	17.9
9	9	Adelie	Biscoe	37.9	18.6
10	10	Adelie	Biscoe	40.5	18.9
11	11	Adelie	Biscoe	39.6	17.7
12	12	Adelie	Biscoe	40.1	18.9
13	13	Adelie	Biscoe	35	17.9
14	14	Adelie	Biscoe	42	19.5

These are good to know about (especially if you need to trim text!), but the real power lies in the next options.

COLUMN QUALITY



☑ Monospaced ☑ Column distribution

☑ Show whitespace □ Column profile

☑ Column quality

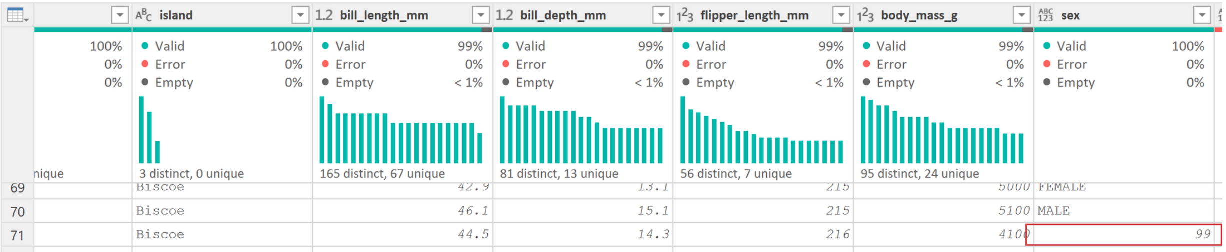
Data Preview

"Column quality" and "Column distribution"

Next, check these two options. You will see a box appear above each column showing what percentage of values are valid, error and empty, along with a distribution of values.

What is a valid cell?

By "valid," Excel simply means it's not empty and it's not an error. That means you could have **nonsensical data in a column and Power Query would still pick it up as "valid,"** such as the value in row 71 of the *sex* column:



		island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
	100%	● Valid 100%	● Valid 99%	● Valid 99%	● Valid 99%	● Valid 99%	● Valid 100%
	0%	● Error 0%	● Error 0%	● Error 0%	● Error 0%	● Error 0%	● Error 0%
	0%	● Empty 0%	● Empty < 1%	● Empty < 1%	● Empty < 1%	● Empty < 1%	● Empty 0%
	nique	3 distinct, 0 unique	165 distinct, 67 unique	81 distinct, 13 unique	56 distinct, 7 unique	95 distinct, 24 unique	
69		Biscoe	42.9	13.1	215	5000	FEMALE
70		Biscoe	46.1	15.1	215	5100	MALE
71		Biscoe	44.5	14.3	216	4100	99

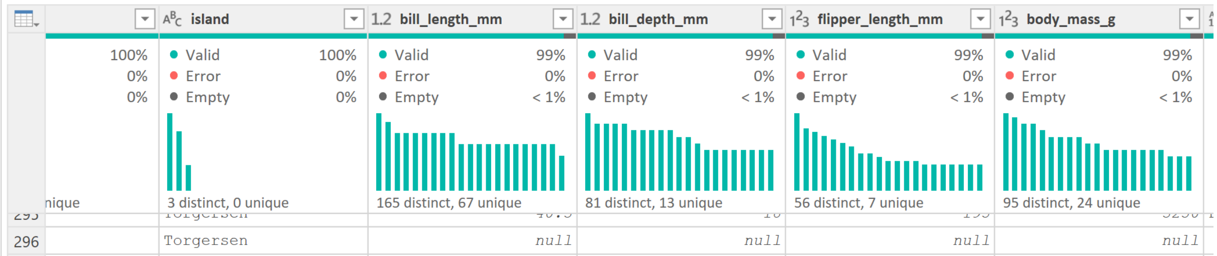
Empty cells

Clearly, "99" is not *really* a valid entry for the sex column. In fact, due to a transcription error these cells were filled with 99 instead of left missing, or *null*.

To see what a true missing value looks like in Power Query, head to row 296.

COLUMN QUALITY

***Power Query's use of* null**



		island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
	100% 0% 0%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 99% ● Error 0% ● Empty < 1%	● Valid 99% ● Error 0% ● Empty < 1%	● Valid 99% ● Error 0% ● Empty < 1%	● Valid 99% ● Error 0% ● Empty < 1%
	nique	3 distinct, 0 unique	165 distinct, 67 unique	81 distinct, 13 unique	56 distinct, 7 unique	95 distinct, 24 unique
296		Torgersen	null	null	null	null

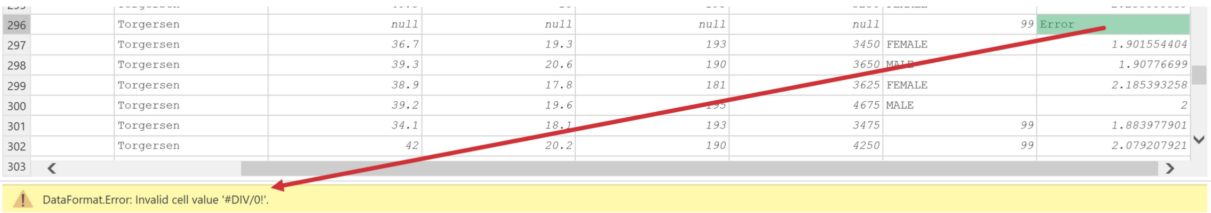
You will see several entries in row 296 marked as *null;* this is the *right* way to leave a cell empty for Power Query! Notice that for each of these columns, <1% of cells are marked as "Empty" in the Column quality menu.

Finally, to understand the Error category, stay on row 296 but check out the *bl_bd_ratio* column. This column was derived in Excel by dividing *bill_length_mm* by *bill_depth_mm*. Because each cell was left empty in Excel, the formula results in a #DIV/0 error.

COLUMN QUALITY

Cell Errors

You can confirm by clicking on the white space outside the *Error* message in this cell:

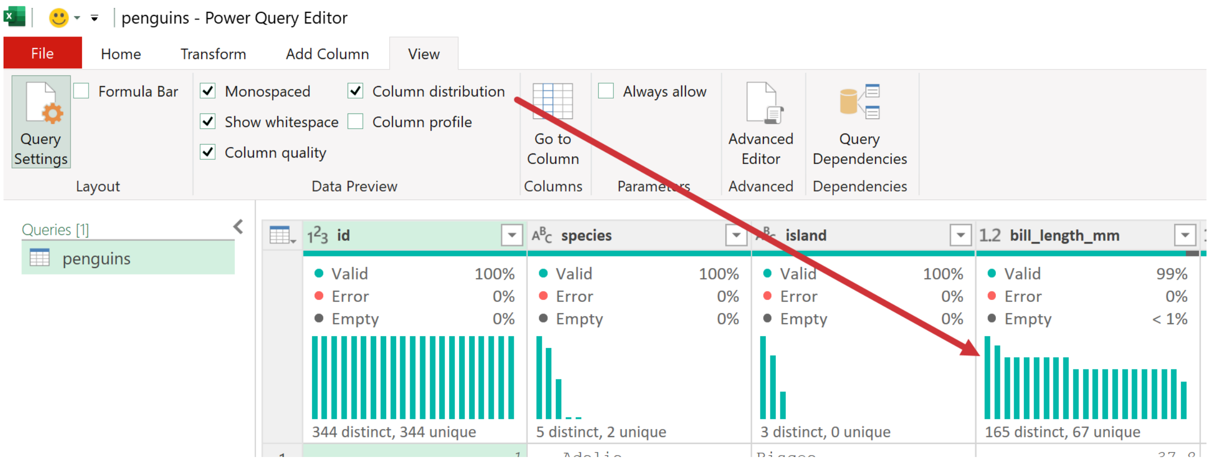


296	Torgersen	null	null	null	null	99	Error
297	Torgersen	36.7	19.3	193	3450	FEMALE	1.901554404
298	Torgersen	39.3	20.6	190	3650	MALE	1.90776699
299	Torgersen	38.9	17.8	181	3625	FEMALE	2.185393258
300	Torgersen	39.2	19.6	195	4675	MALE	2
301	Torgersen	34.1	18.1	193	3475	99	1.883977901
302	Torgersen	42	20.2	190	4250	99	2.079207921
303							

DataFormat.Error: Invalid cell value '#DIV/0!'.

The Column quality menu is a great way to get a quick breakdown of the contents of each column in your data.

Column distribution



☐ Formula Bar
☑ Monospaced
☑ Show whitespace
☑ Column quality
☑ Column distribution
☐ Column profile

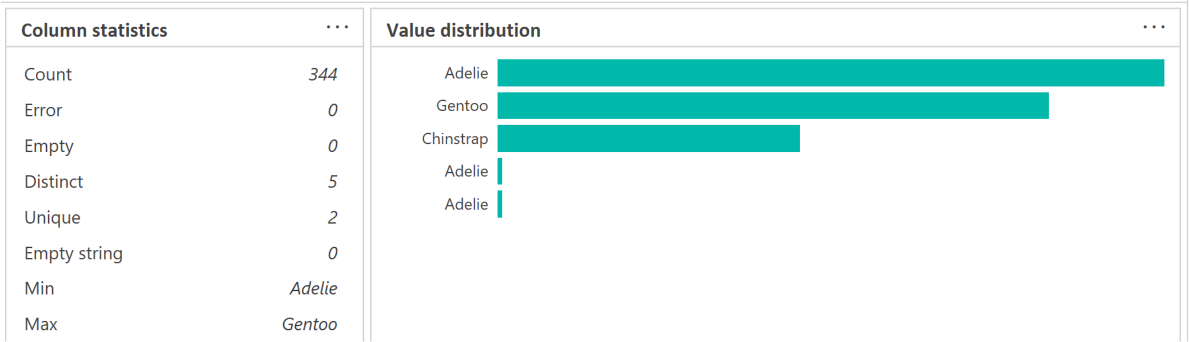
id	species	island	bill_length_mm
Valid 100%	Valid 100%	Valid 100%	Valid 99%
Error 0%	Error 0%	Error 0%	Error 0%
Empty 0%	Empty 0%	Empty 0%	Empty < 1%
344 distinct, 344 unique	5 distinct, 2 unique	3 distinct, 0 unique	165 distinct, 67 unique

This option displays a visualization of the data's distribution along with some other information alongside Column Quality. However, these features are all available in the Column Profile menu, so we'll focus there next.

COLUMN PROFILE

"Column profile"

Last but not least, click on the Column Profile menu, then select any column. Here's what you will see, for example, under the *species* column:



Column statistics

Count	*344*
Error	*0*
Empty	*0*
Distinct	*5*
Unique	*2*
Empty string	*0*
Min	*Adelie*
Max	*Gentoo*

Value distribution

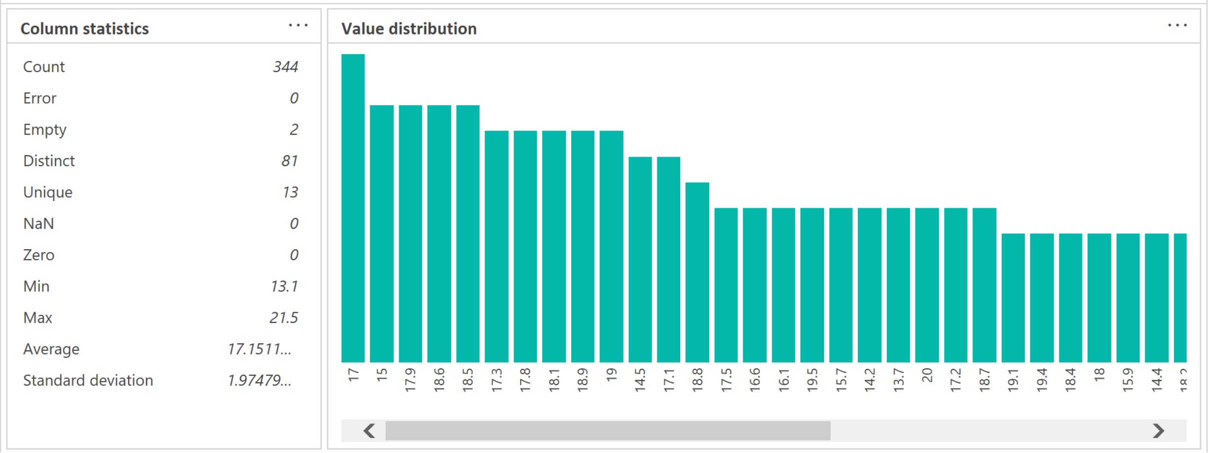
Here we'll get an in-depth breakdown of the values in this column, including a visualization of how many observations are found for each value.

Why three options for Adelie? This has to do with the *extra whitespace* in some of the entries for this category! Once the data is *cleaned* in Power Query, this will go away. But we are profiling first.

COLUMN PROFILE

Column profiling and descriptive statistics

If you click on a quantitative column such as *bill_depth_mm*, you will see some additional column statistics such as the mean and standard deviation alongside the cell counts:



Column statistics

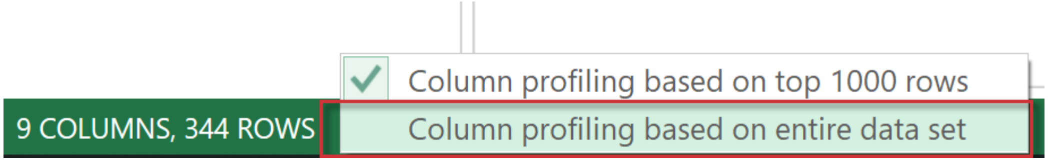
Count	344
Error	0
Empty	2
Distinct	81
Unique	13
NaN	0
Zero	0
Min	13.1
Max	21.5
Average	17.1511...
Standard deviation	1.97479...

Pulling a column's descriptive statistics is a great way to profile it, and Power Query gives us a full suite of measures without writing any formulas.

PROFILING > 1000 ROWS

Overriding the thousand-row limit

If you are working with a dataset of greater than 1,000 rows (not uncommon these days), make sure include *all* of it in the data profiling by clicking toward the bottom of the editor and selecting "Column profiling based on entire data set."


✓ Column profiling based on top 1000 rows
9 COLUMNS, 344 ROWS
Column profiling based on entire data set

THANK YOU

Next steps and resources

With Power Query's data profiling capabilities we were able to:

- Quickly spot incorrectly formatted values
- Determine which columns contain invalid cells
- Visualize the distribution of each variable

Knowing what's going on with the data at this high levels makes you much better prepared to clean it... which of course can be done in Power Query.

To learn more about Excel Power Query can help you for data analysis, cleaning and reporting:

- Subscribe to my newsletter
- Follow me on LinkedIn
- Partner for corporate training on Power Query

What questions do you have about Power Query or making the most of Excel for data analysis? Drop me a line.