Free Alternatives to Excel for Data Cleaning

- Posted by Lee Baker on July 15, 2016 at 6:19am
- View Blog

Pretty much every data rookie starts with Excel. It is a wonderful program for storing, cleaning and analysing (yes, you read that correctly) your data.

Strictly speaking, Excel isn't free, but really – who pays for it these days? If you buy a Windows PC or laptop it'll usually come pre-installed, and if you get a new PC at work your employer will have it pre-installed for you. If you're prepared to look the other way, there are guys who know guys who can get you a copy that fell off the back of a lorry, but I wouldn't endorse that. That would be wrong.

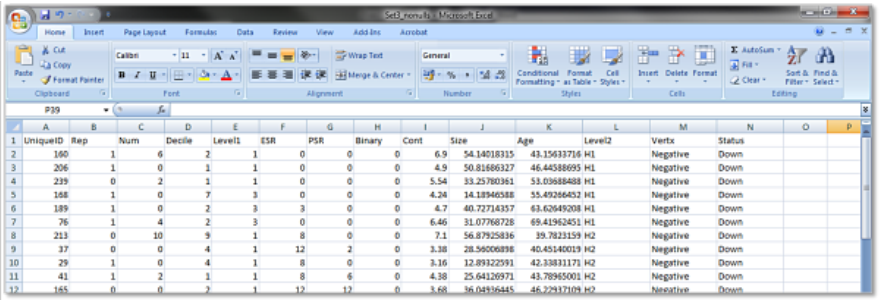
While Excel is a great place to start, once you get into it you quickly realise just how long it takes to clean your data. Worse still, Excel is the source of a great deal of that wasted time!

Never mind – there is a new generation of free data cleaning programs for non-programmers that promise to clean your data quicker, easier and with a lot less hassle, and I'll introduce those that I know of here, namely OpenRefine, Trifacta Wrangler and DataKleenr.

Microsoft Excel

Importing data into Excel is extremely simple, and supports import from Microsoft Access, from the web, text sources, an SQL server, XML, CSV and – thanks to the Data Connection Wizard – a whole host of other data sources. Perhaps the most-used Excel import function though is your fingers. Excel is particularly good with manual data entry and copy/paste operations, which make it the most used program for storing, cleaning and analysing data.

Excel has lots of functions that can help with your data cleaning, such as Remove Duplicates, Find and Replace, Spell Checker, and loads more formulae, such as TRIM(), CLEAN() and SUBSTITUTE().



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	UniqueID	Rep	Num	Decile	Level1	ESR	PSR	Binary	Cont	Size	Age	Level2	Vertx	Status		
2	160	1	6	2	1	0	0	0	6.9	54.14018315	43.15633716	H1	Negative	Down		
3	206	1	0	1	1	0	0	0	4.9	50.81686327	46.44588695	H1	Negative	Down		
4	239	0	2	1	1	0	0	0	5.54	33.25780361	53.03688488	H1	Negative	Down		
5	168	1	0	7	3	0	0	0	4.24	14.18946588	55.49266452	H1	Negative	Down		
6	189	1	0	2	3	3	0	0	4.7	40.72714357	63.62649208	H1	Negative	Down		
7	76	1	4	2	3	0	0	0	6.46	31.07768728	69.41962451	H1	Negative	Down		
8	213	0	10	9	1	8	0	0	7.1	56.87925836	39.7823159	H2	Negative	Down		
9	37	0	0	4	1	12	2	0	3.38	28.56006898	40.45140019	H2	Negative	Down		
10	29	1	0	4	1	8	0	0	3.16	12.89322591	42.33831171	H2	Negative	Down		
11	41	1	2	1	1	8	6	0	4.38	25.64126971	43.78965001	H2	Negative	Down		
12	165	0	0	2	1	12	12	0	3.68	36.04936445	46.22917109	H2	Negative	Down		

On the other hand, Excel has some severe disadvantages. Trailing and leading spaces can be difficult to spot and, although there are formulae to remove them, the procedure to do so can be awkward and quite lengthy:

1. Copy your chosen column to a new sheet
2. Use TRIM() or CLEAN() on the first cell
3. Drag the cleaning procedure to the bottom of the data column
4. Copy your cleaned data to the clipboard and Paste As Values into a separate column
5. Copy the cleaned data back to the original worksheet
6. Repeat for all other columns in your spreadsheet

This is a common problem with cleaning your data in Excel – functions and formulae, although not technically difficult to use, could have been implemented with more style by Microsoft. Would it have been so difficult to have created a button that with a single click would have removed all the trailing and leading spaces from selected columns or the whole worksheet?

The more advanced user could of course build macros and create custom buttons, but they shouldn't have to. Trailing and leading spaces are such a common data cleaning problem that the Microsoft Excel team should have dealt with it already.

Another issue is that some Excel functions operate on selected data, whereas others act on the whole worksheet. If you select a column of data and use Find to identify certain characters, it will identify only those characters in your chosen column. If you now use Replace it will change all such characters in the entire worksheet – which is probably not what you wanted to do, and you may have unwittingly introduced new errors into your data without being aware of it.

The safest way to clean your data in Excel is to copy an individual column to a separate worksheet, perform all your cleaning operations in isolation until you're happy with the result, then copy your cleaned data to your original sheet (or better still, to a new sheet that stores only clean data). The repeated use of Copy, Paste and using multiple worksheets to clean your data can become extremely messy.

Summary

All in all, I would heartily recommend Excel to data novices. It is a very powerful program, is easy to work with and easy to share your work with others. It also has extensive support, forums, books, email courses and all sorts of other help, so if you get stuck with something it's easy to get support. On the other hand it won't take you long before you start to get frustrated at just how much time you waste in cleaning your data by adjusting to Excel's idiosyncracies.

Fear not, though, for there are alternatives to Excel. Read on, dear reader…

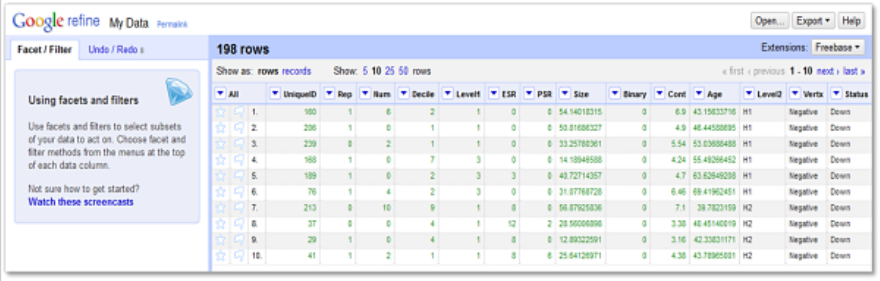
OpenRefine

OpenRefine (previously Google Refine) has the reputation of being 'Excel on steroids', and is a powerful data cleaning tool for text and numerical data that uses your web browser as an interface.

A typical OpenRefine project consists of:

1. Importing your data
2. Transforming it
3. Exporting the result

Importing your data to OpenRefine is fairly simple and leads to the creation of a table of records where each column has a title and each row is numbered. Many data formats are supported for both import and export, including various CSV formats, Excel spreadsheets, JSON, XML and XHTML.



Once imported, there are lots of tools and features offered by OpenRefine to work on your data. All procedures are recorded, making it possible to browse and undo at any time, which often proves extremely useful.

Most operations on data are row-based, column-based, or cell-based.

Row-based operations are limited to marking and deleting selected rows, and frustratingly there is no Add Rows functionality. Filtered and facetted searches are a powerful means to explore, search, and edit data, allowing rows to be selected by listing the distinct values in a column and the number of instances

Summary

OpenRefine, although it has a steeper learning curve than Excel, is much more powerful. Just by using the common transformations you can trim hours off your data cleaning, and the powerful undo/redo functionality take away a lot of the stress compared with Excel.

Although OpenRefine has a lot of well-thought out features, I find it quite annoying that a number of common, simple operations on data are un-necessarily complicated, requiring more mouse clicks than it really needs to. I've also found that lengthy data cleaning sessions lead to a degradation in performance, leading to either crashes or a need to save, close and re-open.

OpenRefine doesn't have extensive support, although there are a few How-To videos and a book. If you get stuck and don't know where to get help you may be out of luck.

All in all, OpenRefine is powerful, can save time and stress, but can also increase stress in some circumstances.

I think it might be fair to say that OpenRefine is like Marmite – you'll either love it or hate it!

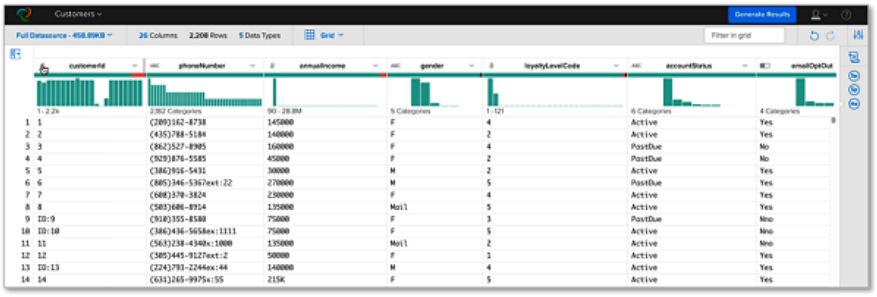
Check out OpenRefine here

Trifacta Wrangler

Trifacta Wrangler (previously Data Wrangler), unlike Excel and OpenRefine, is a semi-automated data cleaning tool for a variety of data types, including text and numerical data, binary data and several others.

When you download and install the application you're required to create an account and log in each time you use it. Apparently all your data wrangling operations and workflows are done locally in your machine, and the account creation process is to help you keep the program up-to-date.

On opening the application and loading a dataset (supported formats: CSV, TSV, JSON, Excel), Trifacta Wrangler tries to figure out the data types of each column, then gives a visual summary of your dataset. Helpfully, it gives a horizontally-stacked bar chart at the top of each column to show the proportion of valid, mismatched and missing values, and this is a great help in identifying errors.



Data cleaning with Trifacta Wrangler is manual and column-based, but is very much easier to use than Excel or OpenRefine. Identify an erroneous data pattern in a given column and Trifacta Wrangler will automatically apply your data cleaning operation throughout the column, although you have to manually identify the errors in your data and provide your own solution to the issues. You can split and merge columns easily and it also has other, more complex features that can be accessed via the Transformer.

Summary

The focus of Trifacta Wrangler is firmly on business data, and on preparing your data for porting to Tableau. If you need help, it has a built-in tutorial, but little other support.

Trifacta Wrangler is a very impressive program that is visually pleasing and easy to work with. Its advantage over Excel and OpenRefine is in its semi-automation, allowing you to slash your data cleaning operations to a fraction of the typical time taken. If OpenRefine is not to your taste, Trifacta Wrangler just might be.

Check out Trifacta Wrangler here

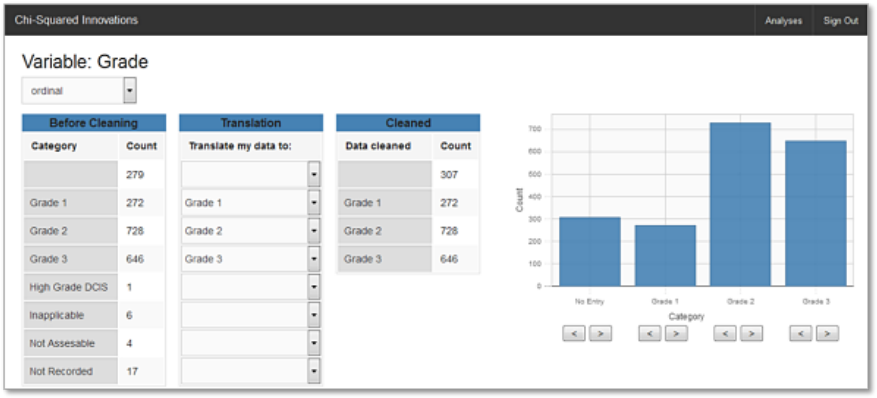
DataKleenr

DataKleenr is the newest entrant into the free data cleaning software space, and is a fully-automated solution for cleaning text, numerical and binary data.

It is cloud-based, so there's no need to download or install anything, and all you need is a browser and an internet connection. You need to create an account and log in each time, and all your data cleaning operations are performed on the cloud, encrypted and saved to your private personal workspace, where you can return to your projects at any time.

When you load a dataset (currently, only comma-separated CSV files are supported), DataKleenr automatically detects the type of data for each column and assigns it a label of Continuous, Ordinal or Nominal, and gives a visual stacked bar chart summary of your data showing the proportion of cleaned, uncleaned and missing values. You can change any of the data type assignments via a simple dropdown menu.

Data cleaning is column-based in DataKleenr, but is mostly already done at the upload stage. It has intelligent algorithms that automatically decide how to clean your data, so mostly you will only need to check the final result. Click on a variable to check the details for an individual variable.



For ordinal and nominal data, DataKleenr decides on the major families contained in each variable, corrects any spelling errors, and excludes data that it considers to be invalid. Bar charts help to show the distribution and order of categories. All decisions can be countermanded simply and easily, and DataKleenr learns from this, improving performance of future data cleaning operations.

For continuous data, invalid entries, such as text, symbols or mixed data types are automatically purged. Statistical outliers are identified and visualised with a Box-and-Whiskers plot and can be excluded individually, en-masse or not at all.

When you're finished, you can download your data as a comma-separated CSV file.

Summary

DataKleenr has advantages over other programs in that it is fully automated, intuitive and simple, typically completing data cleaning in a few minutes. On the other hand, DataKleenr is not a tool for all types of data, nor is it meant to be. The focus is firmly on science-type data, on doing the simple things well and on providing visual aids at all stages of data cleaning to produce data that is analysis ready quickly and easily.

If you need help with DataKleenr there are few resources (not surprising since it's so new), but there is an accompanying book and you can contact the company directly who will help with any issues you might have.

Check out DataKleenr here

Conclusions

While there are many data cleaning programs in the commercial space, few of them are offered for free. Some offer free time-limited trials or demos, while others offer free programs with limited functionality. There are also free data cleaning plugins for use in programs like R, but you need to be a confident programmer to be able to use them.

For non-programmers looking to clean their data for free there are few tools available, and I've reviewed those that I know of above.

Microsoft Excel is the obvious starting point for pretty much everybody, but it has its disadvantages. The next generation of tools, such as OpenRefine, Trifacta Wrangler and DataKleenr are quicker, easier and more accurate options for those wishing to be more efficient in their data cleaning.

What do you think?

Do you know of any other free data cleaning programs for non-programmers?

What are your experiences with these programs?

Join the debate below and let me know your thoughts...

About the Author

Lee Baker is an award-winning software creator with a passion for turning data into a story.

A proud Yorkshireman, he now lives by the sparkling shores of the East Coast of Scotland. Physicist, statistician and programmer, child of the flower-power psychedelic '60s, it's amazing he turned out so normal!

Turning his back on a promising academic career to do something more satisfying, as the CEO and co-founder of Chi-Squared Innovations he now works double the hours for half the pay and 10 times the stress - but 100 times the fun!

PS - Don't forget to connect with me in Twitter: @eelrekab

Other DSC Articles by the same Author

- Why Good Data Scientists are Worth the Big Bucks
- 50 Shades of Grey ... The Psychology of a Data Scientist

Like
12 members like this

Share Tweet Facebook

Like 0

- < Previous Post
- Next Post >

Comment

You need to be a member of Data Science Central to add comments!

Join Data Science Central

 Comment by Lee Baker on May 22, 2018 at 2:21am

@Edwin

thanks for that - when you've evaluated them, come back and let us all know how you got on. I'm sure your experiences would be valuable to other readers!

Cheers

 Comment by Edwin Chuy on May 21, 2018 at 12:52pm

There's also Talend Free Data Preparation which is probably a relative basic tool. I'll give those mentioned in the article a try to see how they work for me.

 Comment by Lee Baker on May 30, 2017 at 1:14am

@Venu

It's not enough that you spam me on my own site, but now you feel the need to do it here at DSC.

It's not appreciated!

Your comments have been deleted at both sites. Continue and you will find yourself blocked permanently.

For everyone else, please continue commenting - I love your positivity and look forward to reading more of your thoughts and suggestions.

 Comment by Maksymilian Piechanowski on July 26, 2016 at 11:27pm

Thank you!

 Comment by Gary Revell on July 25, 2016 at 8:20am

Hi,

A while ago I used Qlikview for loading CSV and flat text files, and then massage the data to create new views which you could then write to new files for processing.

The scripting language was pretty good too for wrangling, and you could setup keys using same column names on different tables to join data. Then you can filter data in the UI before writing it out.

I used it for various payroll projects where I wanted to cross reference employee data that was sent in different data files to the payroll vendor.

Best regards

Gary

 Comment by Lee Baker on July 21, 2016 at 1:39pm

@Rusul

I agree - Excel is a great program, particularly in teaching students. Despite having alternatives at my fingertips, I still use Excel a lot.

 Comment by rusul issam mahdi on July 21, 2016 at 11:30am


Comment by Lee Baker on July 20, 2016 at 10:58am

@Rick

In a previous life I was a consultant medical statistician where I was responsible for managing the datasets of hundreds of doctors, surgeons, nurses, pathologists, micro biologists, etc., and in the 7 years that I did the job, not a single one of them could program in Excel, create macros, scripts or anything else like that.

They all cleaned their data by eye or - worse still - gave me the unclean dataset and expected me to do it for them.

This is the crux of the problem. Most of the people that have to work with data can't program, and shouldn't have to, and this is why the new generation of data cleaning programs are so exciting.


Comment by Rick Henderson on July 20, 2016 at 10:26am

Interestingly enough, the source for the formula Eric describes is an Excel help document related to the CLEAN function: Remove spaces and nonprinting characters from text


Comment by Rick Henderson on July 20, 2016 at 10:09am

This is a great resource for non-programmers as stated, but as you describe yourself as a statistician, physicist, and programmer I'm surprised you didn't mention that most of the problems you ascribe to using Excel are eliminated if you write subroutines/scripts/macros in VBA.

- ‹ Previous
- 1
- 2
- Next ›
- Page 2 Go

RSS

VIDEOS

- DSC Webinar Series: Clean Data & Accurate ML Models

  Added by Tim Matteson 0 Comments 1 Like

- DSC Webinar Series: Analyze Data Faster with an Open Source Columnar Database

FOLLOW US

@DataScienceCtrl | RSS Feeds