topcaats
audit reinvented

Importing and Cleansing Data in Excel

Contents

Preface 3

1 Introduction 4

1.1 Welcome 4

2 Getting your data into Excel 5

2.1 File types 5

3 Importing Data - worked example 7

3.1 Text Import Wizard – Step 1 of 3 7

3.2 Text Import Wizard – Step 2 of 3 8

3.3 Text Import Wizard – Step 3 of 3 10

4 Cleaning the data – worked example 12

4.1 Understanding where you need to get to 12

4.2 Understanding what you're working with 12

4.3 Setting up column headers 13

4.4 Adding additional columns 14

4.5 Applying AutoFilters 21

4.6 Checking data 21

4.7 Deleting unwanted columns 21

4.8 Congratulations 22

5 Formulas 23

5.1 IF statements 23

5.2 AND and OR 24

5.3 LEFT, RIGHT and MID 24

5.4 TRIM 25

5.5 VALUE 25

6 Techniques 26

6.1 AutoFilter 26

6.2 AutoFilter in combination with a filter column 27

6.3 Fill down formulas – dealing with subheadings and grouped data 27

6.4 Joining cells 28

6.5 Multiple row headers / record split across 2 rows 28

6.6 Convert negative to positive 28
6.7 Move “-“ from right to left 29
6.8 Find and replace 29
6.9 Deleting blank rows 29
6.10 Deleting subtotals / subheadings 30

Preface

This eBook has been written by Reinvent Data Limited to help our users import data into Excel, and appropriately clean that data so that it can be analysed using data analysis tools including TopCAATs.

Whilst this eBook is not complicated and advanced Excel skills are not necessary, a basic understanding of Excel is assumed.

The example file and video walkthrough that accompany this eBook can be downloaded from the TopCAATs website at www.topcaats.com/etc/?mod=ebook.

If you are printing this eBook, please use double sided printing where available to help save paper!

If you have any questions relating to this eBook, or would like to find out more about TopCAATs, please email us at enquiries@reinventdata.com

1 Introduction

1.1 Welcome

Welcome to this TopCAATs eBook on importing and cleansing data. Sometimes one of the most challenging parts of data analysis is simply getting your data into the data analysis software you're using. Thankfully, many modern accounting systems are now able to export data directly to Excel, usually in a tabular format that is ready for data analysis. However, some computer systems are not designed for outputting to spreadsheets, or generate reports that are not analysis-friendly.

This eBook will guide you through some really useful (and clever) techniques to enable you to quickly and easily get your data into the format that you need it in.

The eBook is structured around a worked example which is covered in sections 3 and 4. The example text file is a sales listing (which you can download here www.topcaats.com/etc/?mod=ebook) and presents a number of challenges. It is somewhat of a worst-case scenario of a system report and has been designed to incorporate many of the challenges you will face with real life system reports.

Whilst the example data is very short, the techniques that will be demonstrated are exactly the same as you would use for a report with thousands of records, and the time taken to process a larger report would be exactly the same, as we use existing characteristics within the data rather than looking at each data row individually.

The best way to use this eBook is to work through the example as you read the eBook. If you need additional guidance there is a video walkthrough available here www.topcaats.com/etc/?mod=ebook. Working through the example and practicing the techniques is the easiest way for you to learn and remember them.

At the end of the eBook are sections covering the formulae and techniques used, and provide further detail if you need it. These are also a good reference for the future.

You do not need to have TopCAATs installed to work through the example. However, if you haven't already tried out TopCAATs we recommend downloading the free 30 day trial from www.topcaats.com, as you'll probably find it saves you several hours a week!

TopCAATs is a data analysis add-in for Excel with over 100 tools. Although many of the tools are focused towards accountants and auditors, anyone using Excel on a regular basis is likely to benefit from having TopCAATs installed!

2 Getting your data into Excel

2.1 File types

The way you get your data into Excel will depend on the file type, which can be determined by the file extension (3 letters that follow the file name). Some systems have a choice of export options. For import into Excel the most desirable formats are Excel files (.xls or .xlsx) or comma separated values (.csv).

Here are some of the file types that you may encounter:

2.1.1 Excel files pre-office 2007 (.xls)

These can be opened directly in all versions of Excel using File > Open, or double clicking the file in Windows Explorer.

2.1.2 Excel 2007 files (.xlsx)

These files will open automatically in Excel 2007. Previous versions of Excel require a convertor to open these files. If you don't already have it installed Excel can usually download this automatically from the Microsoft website.

2.1.3 Comma-separated values (.csv)

These can be opened directly in all versions of Excel using File > Open, or double clicking the file.

2.1.4 Text files (.txt)

Plain text files can be opened in all versions of Excel using File > Open, however they will need to be imported using the Text Import Wizard, which runs automatically when you open a text file in Excel. The Import Wizard is covered in the worked example that follows (section 3).

2.1.4.1 Delimited text files

Delimited text files use a character to symbolise where the next column begins. A text file delimited with a semicolon (;) may look as follows

```
Salesperson;Invoice Number;Date;Product Number;Product Description
Smith, James;SINV0001;12.12.08;BVV12345;Ballpoint pen (red)
```

To import a delimited text file

- Open the file using File > Open
- Select Delimited and click Next
- Select the delimiter(s) that are used in your file to separate columns and click Next
- Check that the data looks correct in the Data Preview at the bottom of the window
- Set data formats for each column. You can leave most columns as General, but should always identify date columns as such, as this converts them into a date format recognised by Excel.
    - Select the columns containing dates and click the "Date" button.
    - Select the order of days (D), months (M) and years (Y) in the current date format. In the example above the format is "DMY".
- Click Finish to complete the wizard

2.1.4.2 Fixed width text files

Fixed width text files are arranged into columns using spaces to ensure the columns line up.

Please see Chapter 3 for a worked example of importing fixed width text files.

Please note that there may be other information in the file that does not fit neatly into the columns you define (e.g. sub headings or totals). This is rarely a problem and can be fixed later, as we will demonstrate in the worked example.

2.1.5 ODBC

Excel has the native capability to import data directly from a database. This process is beyond the scope of this eBook.

2.1.6 Other file types

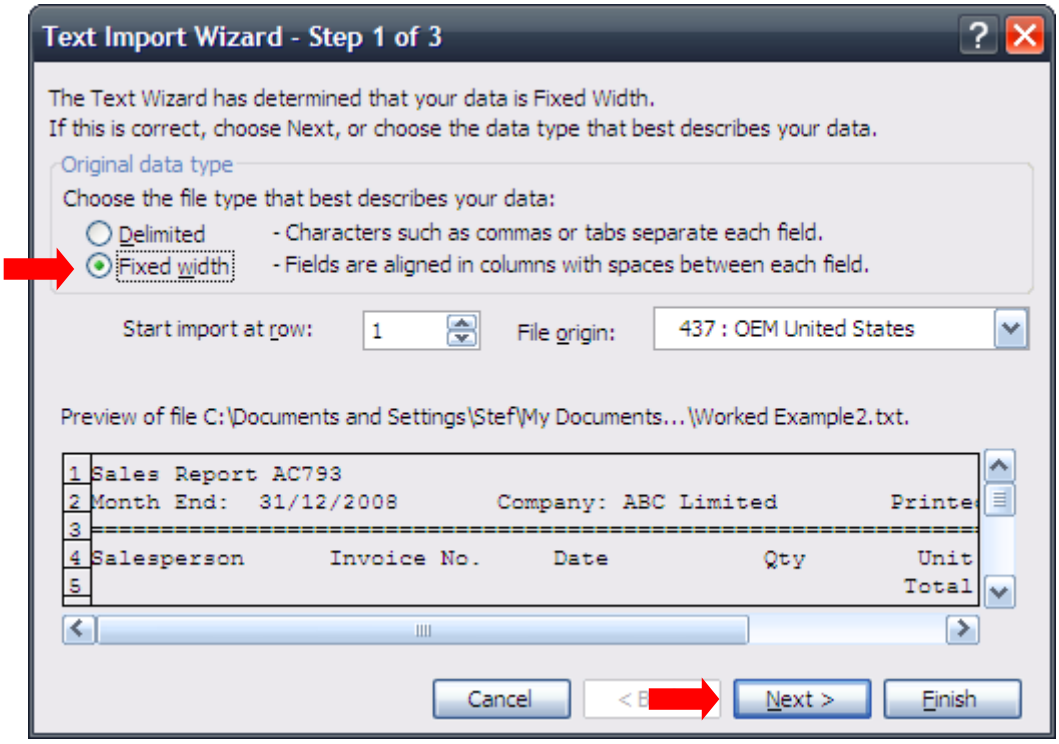
Some other files can be opened in Excel, depending on their nature. Many "proprietary" formats are actually just text files with a different extension. These can often be opened in Excel either directly, or by changing the extension to ".txt" and then opening them as text files. Many "Print files" (.prn) can be opened in this manner.

3 Importing Data - worked example

First, open the example text file (downloadable here, www.topcaats.com/etc/?mod=ebook) in Excel using File > Open and navigate to the location where the file is saved. You may need to click the "Files of Type" box at the bottom of the window and change it to "All Files (*.*)" or "Text Files (*.prn; *.txt; *.csv)" for the file to show up.

3.1 Text Import Wizard – Step 1 of 3

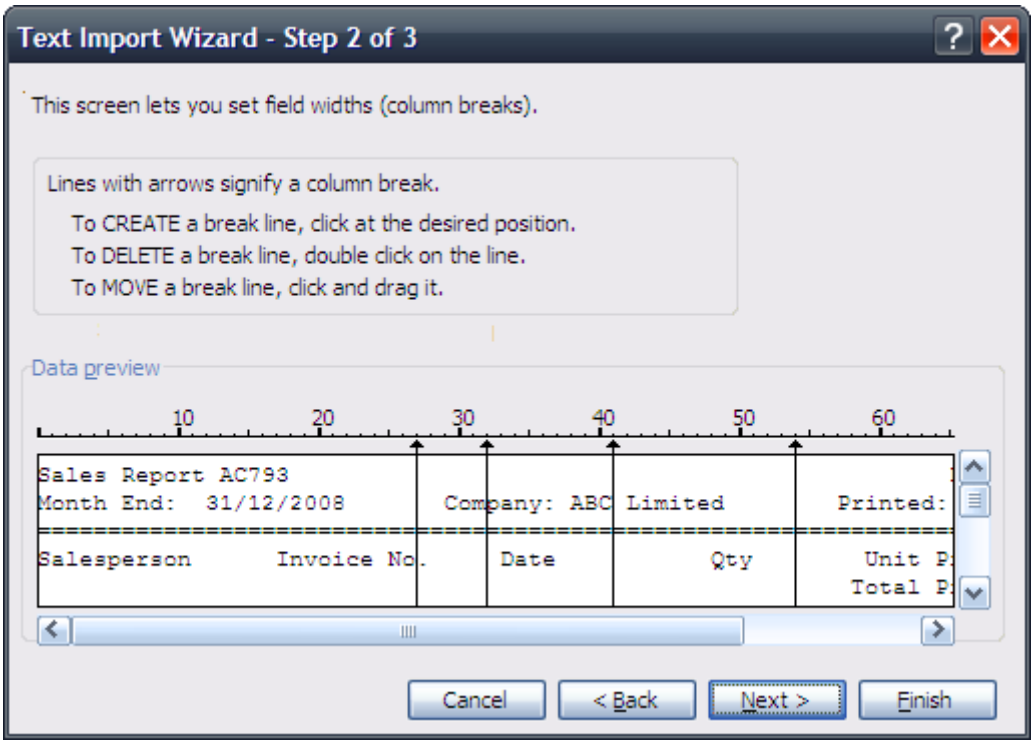
Firstly you need to identify whether the file is Delimited or Fixed Width. The preview at the bottom of the window shows that the text file is arranged neatly into columns already. This indicates that the file is "Fixed Width". Ensure the fixed width button is selected and click Next



3.2 Text Import Wizard – Step 2 of 3

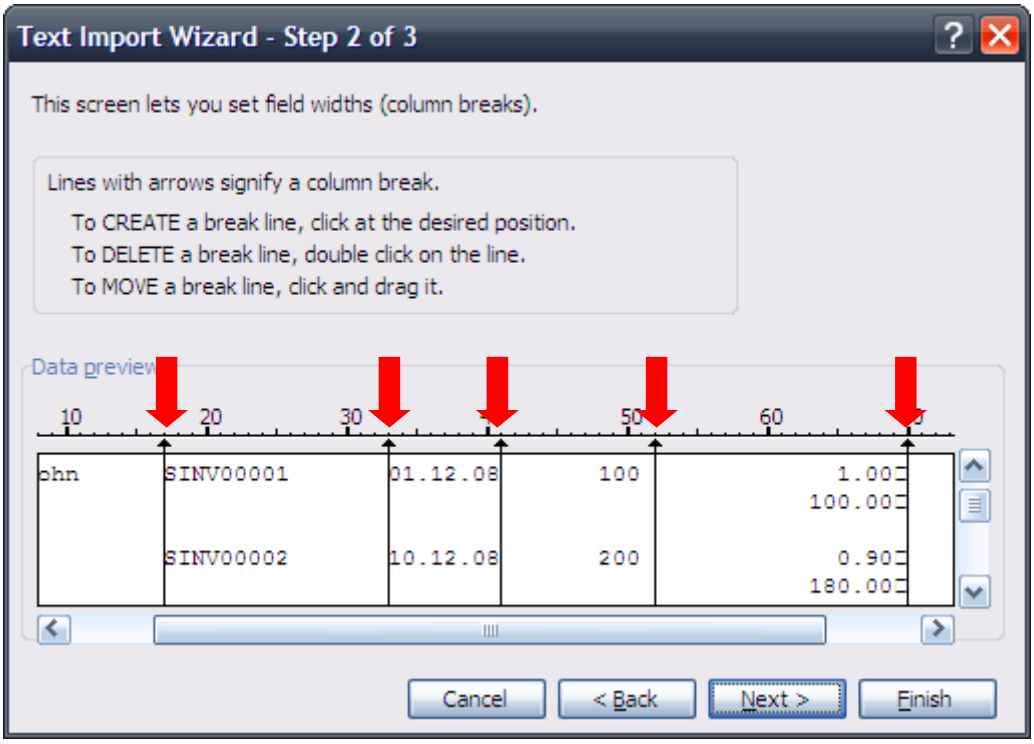
As the file is Fixed Width, you need to tell Excel where you want to split the data into columns. It will have a go at doing this itself, but will usually need to be checked or adjusted.

Looking at the data preview at the bottom of the window we can see that the column splits are not in the correct places. Excel has incorrectly guessed the column splits, and we need to look at the data portion of the file, to work out where they should be. Scroll down in the Data preview window to where the data is arranged into columns.



The screenshot below shows the data portion of the file that we need to base the column breaks on. You next need to move the column breaks as shown in the screenshot by clicking and dragging them in the data preview.

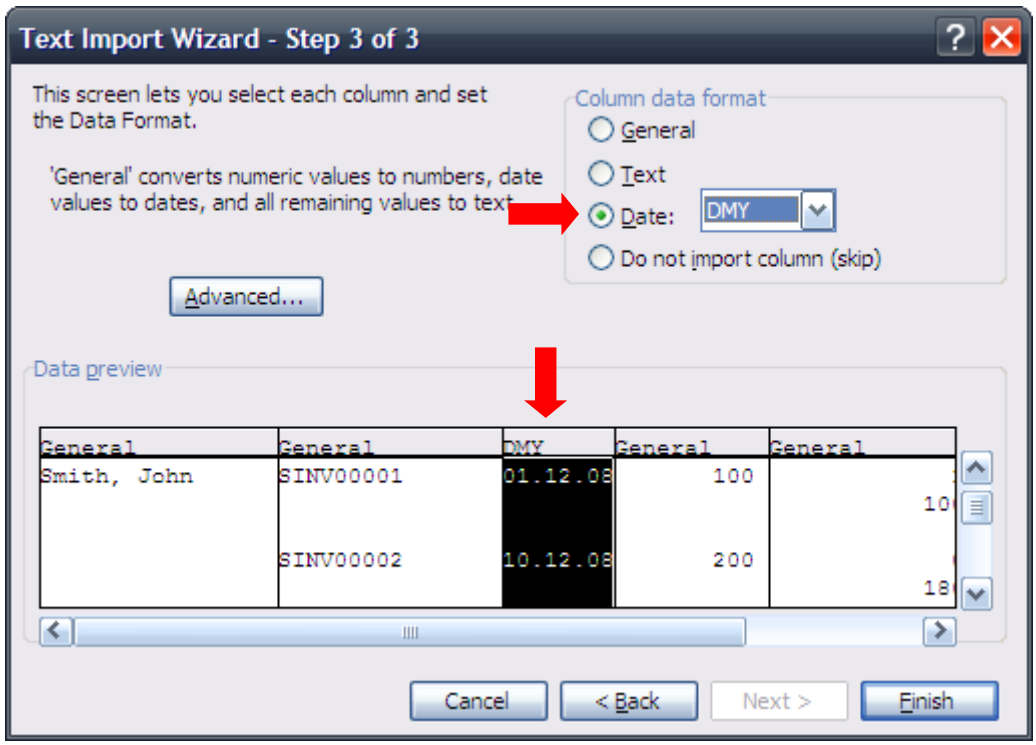
- The first break line should be to the immediate left of the invoice number with the first column containing the salesperson's name (position 17).
- The second break line should be just to the left of the date (position 33).
- The third break line should be just to the right of the date (position 41).
- The fourth break should be one character to the right of the Qty field. This is to ensure the negative sign (at the right of the number in this case) is captured in that field (position 52).
- The fifth break should be one character to the right of the Unit Price / Total Price field, again to capture any negative signs (position 70).
- Remove any other breaks as these are not required.
- Click Next.



3.3 Text Import Wizard – Step 3 of 3

The final step of the wizard requires you to set data format for each column. You can leave most columns as General, but should always identify date columns as such, as this converts them into a date format recognised by Excel.

- Select the columns containing dates and click the "Date" button.
- Select the order of days (D), months (M) and years (Y) in the current date format. In this case the format is "DMY".

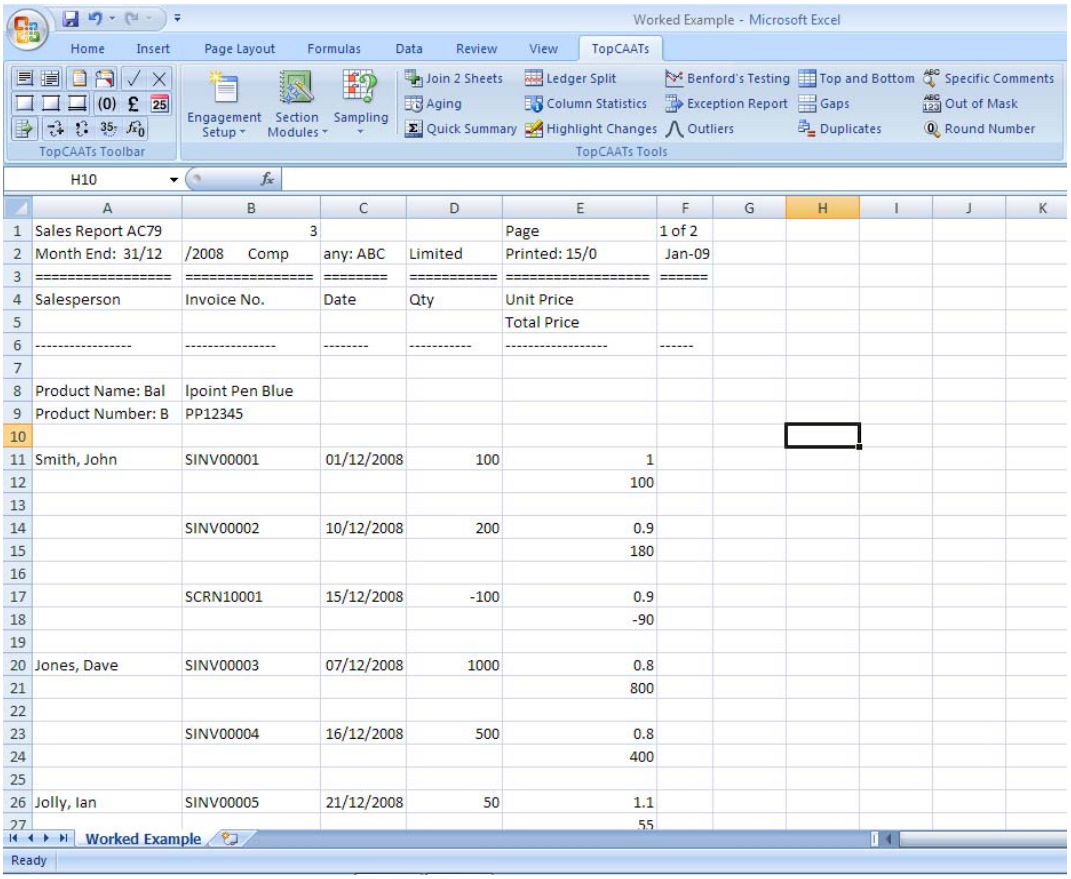


Finally, click "Finish" to complete the wizard

toptip

Sometimes Excel will incorrectly identify a column as dates when you import it. For example, if you have a column with values such as "1/10", "2/10", "3/10", etc. then Excel may translate these to dates. To prevent this you can specify the column as "Text"

Your data should appear as below (after you have resized all columns to fit the contents). Although the data is now in Excel, it is far from being ready to perform data analysis on. This is where data cleansing is needed.



	A	B	C	D	E	F
1	Sales Report AC79	3			Page	1 of 2
2	Month End: 31/12	/2008 Comp	any: ABC	Limited	Printed: 15/0	Jan-09
3	==========	==========	========	==========	==========	======
4	Salesperson	Invoice No.	Date	Qty	Unit Price	
5					Total Price	
6	----------	----------	--------	----------	----------	------
7						
8	Product Name: Bal	lpoint Pen Blue				
9	Product Number: B	PP12345				
10						
11	Smith, John	SINV00001	01/12/2008	100	1	
12					100	
13						
14		SINV00002	10/12/2008	200	0.9	
15					180	
16						
17		SCRN10001	15/12/2008	-100	0.9	
18					-90	
19						
20	Jones, Dave	SINV00003	07/12/2008	1000	0.8	
21					800	
22						
23		SINV00004	16/12/2008	500	0.8	
24					400	
25						
26	Jolly, Ian	SINV00005	21/12/2008	50	1.1	
27					55	

Worked Example

Ready

Note that Excel has automatically identified the negative numbers, even though the minus sign was on the right hand side of the number.

When you import other file types, this may not happen automatically, so you may need to perform this step manually. Please see section 6.7 for details on how to perform this step.

4 Cleaning the data – worked example

The approach to cleaning data should always be the same

1. Understanding where you need to get to
2. Understanding what you're working with
3. Setting up column headers
4. Adding additional columns to ensure each record has all the required information on a single row
5. Applying auto filters and extract visible cells
6. Checking data
7. Deleting unwanted columns

4.1 Understanding where you need to get to

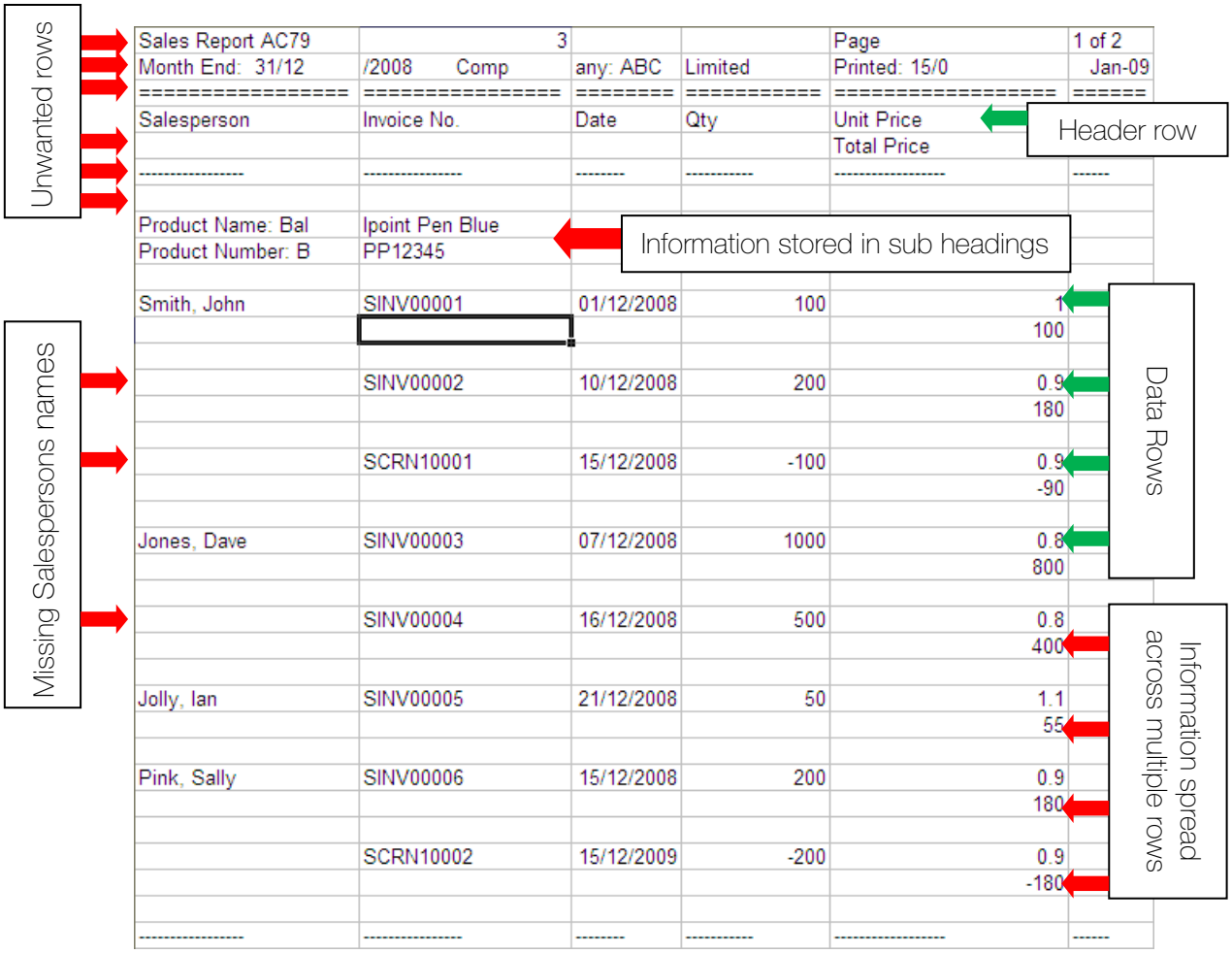
The goal of cleaning data is getting the data into a suitable format for data analysis. The key to this is a rectangular block of data, with a single header row and no blank rows or columns, and no total rows.

Each record needs its own data row, and all the information for that record should be on a single row. It also helps to have an understanding of what information you need for each record (i.e. what columns you need in the final report).

4.2 Understanding what you're working with

When you look at the example file (which has been created to include a wide variety of common issues), you can see the following features

- A variety of unwanted rows, including page headers, duplicated column headers, separator rows, blank rows and total rows
- Information spread across two rows, e.g. each record has 2 rows, with "Total Price" appearing on the second row
- Useful information stored in sub headings, e.g. the product name and number
- Missing information, e.g. the information is grouped by sales person, and the sales persons name appears only once for each product



Sales Report AC79		3			Page	1 of 2
Month End: 31/12	/2008 Comp		any: ABC	Limited	Printed: 15/0	Jan-09
==========	==========		========	==========	==========	=======
Salesperson	Invoice No.		Date	Qty	Unit Price	
					Total Price	
----------	----------		--------	----------	----------	------
Product Name: Bal	Ipoint Pen Blue					
Product Number: B	PP12345					
Smith, John	SINV00001		01/12/2008	100	1	
					100	
	SINV00002		10/12/2008	200	0.9	
					180	
	SCRN10001		15/12/2008	-100	0.9	
					-90	
Jones, Dave	SINV00003		07/12/2008	1000	0.8	
					800	
	SINV00004		16/12/2008	500	0.8	
					400	
Jolly, Ian	SINV00005		21/12/2008	50	1.1	
					55	
Pink, Sally	SINV00006		15/12/2008	200	0.9	
					180	
	SCRN10002		15/12/2009	-200	0.9	
					-180	
----------	----------		--------	----------	----------	------

4.3 Setting up column headers

Before we begin cleaning the data we need to ensure that we have a header row at the top of our data, and every column should have a unique header.

In this example, we can simply delete the first 3 rows, which are page headers and do not contain useful information. This will leave us with suitable column headers in the first row.

Alternatives are inserting a new row and adding your own (or copying) headers, or hiding the first 3 rows.

There is also a missing column header in column F (column F is not a blank column, but it has no header). It's best to add a header now to avoid problems later (e.g. forgetting that this column isn't blank and entering formulas in it).

4.4 Adding additional columns

Each piece of information you want to capture (or field) needs its own column. Many of the required fields will already have their own column with the data correctly on the data rows. In this example we do not need to do anything for the “Invoice Number”, “Date”, “Qty” or “Unit Price”.

The example below guides you through creating additional columns to ensure all the data is captured correctly for our example – “Total Price”, “Sales Person”, “Product Name” and “Product Number”.

The columns you need to add will depend on your data, but the techniques described will be useful for capturing the vast majority of data

The approach is to pick a field that is not in the correct place (i.e. does not currently have its own column, with the data on the correct row) and to work out how to get the right data into that column using Excel’s formulas.

Please note that throughout this process there will be a lot of cells/rows that contain seemingly “nonsense” data. It may look strange and unnerving at first, but these will be filtered out later. If you look at the data rows (coloured green in the video and screenshots below), you will notice that they actually contain the useful information.

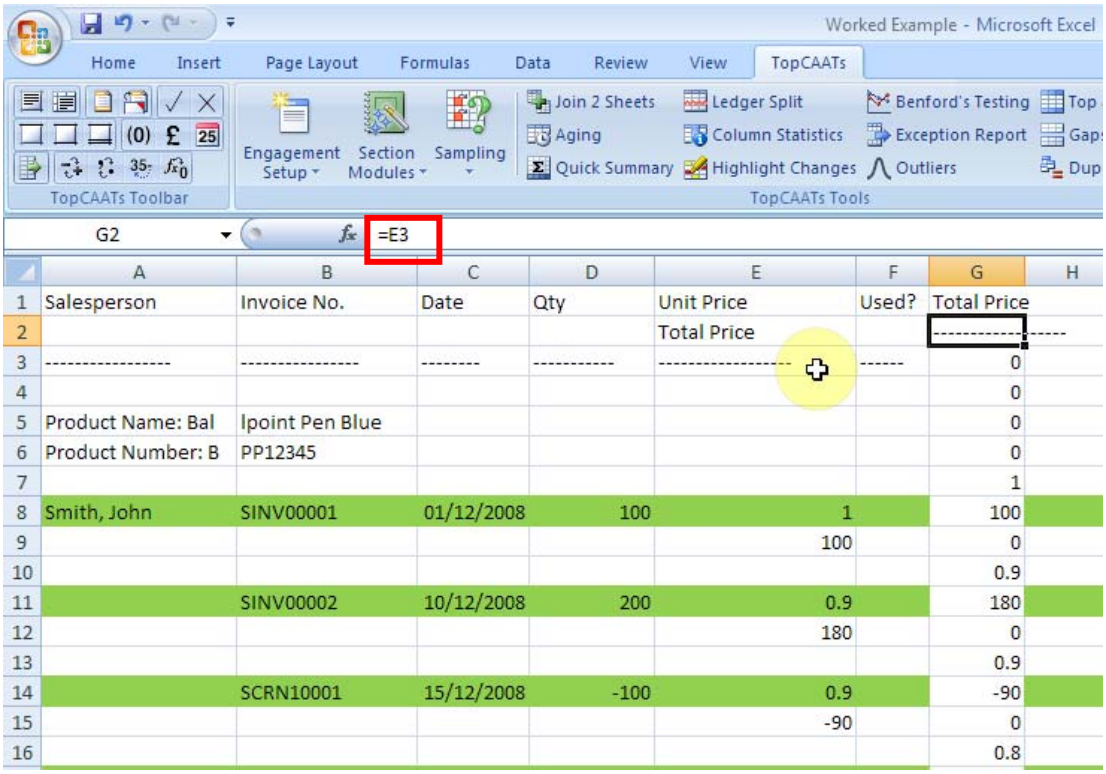
4.4.1 Total price

The total price for each invoice is currently being stored in a separate row, immediately below the main data record i.e. each record is actually spread across 2 rows. To get the total price into its own column on the correct row we can use a very simple "=" formula. In cell G2 enter the formula

=E3 and fill this down to the bottom

How this formula works:

It simply returns the value from the row below in column E – i.e. where the Total Price is currently being stored



	A	B	C	D	E	F	G	H
1	Salesperson	Invoice No.	Date	Qty	Unit Price	Used?	Total Price	
2					Total Price			
3	----------	----------	-------	----------	----------	------	0	
4							0	
5	Product Name: Bal	Ipoint Pen Blue					0	
6	Product Number: B	PP12345					0	
7							1	
8	Smith, John	SINV00001	01/12/2008	100	1		100	
9					100		0	
10							0.9	
11		SINV00002	10/12/2008	200	0.9		180	
12					180		0	
13							0.9	
14		SCRN10001	15/12/2008	-100	0.9		-90	
15					-90		0	
16							0.8	

4.4.2 Salesperson

In this example the salesperson is only listed once for each product, so the majority of data rows do not include any salesperson's name. We need to use what is known as a "fill down" formula.

Fill down formulas look at a certain column for specific criteria, and if that criteria is met they return the value in that column, if not they return the value in the cell above, in the current column.

Here we can see that the salesperson is in the format "Surname, First name", so a suitable criteria to match would be "*,*"

In cell H2 enter the formula

=IF(ISERROR(SEARCH("*,*",A2)),H1,A2) and fill this formula down

How this formula works:

Let's break this formula down into its constituent functions

=IF(ISERROR(SEARCH("*,*",A2)) , H1, A2)

Syntax =IF(*test*, *result if true*, *result if false*)

This is an IF formula, with a test and a result for true and false. If the test returns "True" then the result will be H1 (i.e. the cell above the current cell), if the test returns "False" then the cell will be A2 (i.e. the cell containing the sales person).

Let's look at the test =(ISERROR(SEARCH("*,*",A2))

Syntax =ISERROR(*value*)

This will return True if the value returns an error, and False if the value does not return an error.

Lets now look at the Search function =SEARCH("*,*",A2)

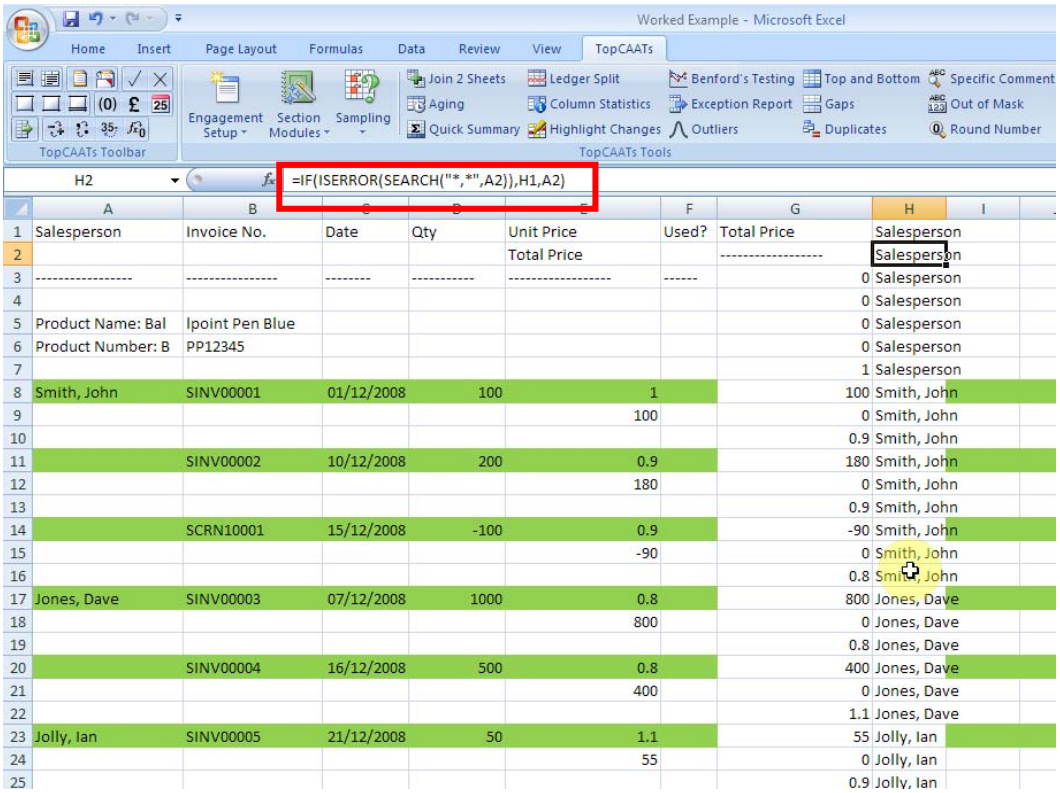
Syntax =SEARCH(*find text*, *within text*)

The first part is the text string we want to find, in this case "*,*", and it must be entered in "quotes" as it is a text value.

The second part is where to look for this text, in this case A2, the cell that might contain the sales person's name.

The search function will return the character position where the text is first found, assuming the text is found. When the text is not found it returns an error.

Putting this all together, this formula works by looking at the value in column A and asking the question 'does it match the pattern "*,*"?' If it does match the pattern, then the ISERROR will return FALSE and the IF will return the sales person's name. If it doesn't match the pattern then the ISERROR will return TRUE and the IF will return the value above (i.e. the last sales person). As we fill this formula down, it will keep returning the same sales person until it finds a new valid sales person in column A.



	A	B	C	D	E	F	G	H
1	Salesperson	Invoice No.	Date	Qty	Unit Price	Used?	Total Price	Salesperson
2					Total Price		----------	Salesperson
3	----------	----------	--------	----------	----------	------	0	Salesperson
4							0	Salesperson
5	Product Name: Bal	Ipoint Pen Blue					0	Salesperson
6	Product Number: B	PP12345					0	Salesperson
7							1	Salesperson
8	Smith, John	SINV00001	01/12/2008	100	1		100	Smith, John
9					100		0	Smith, John
10							0.9	Smith, John
11		SINV00002	10/12/2008	200	0.9		180	Smith, John
12					180		0	Smith, John
13							0.9	Smith, John
14		SCRN10001	15/12/2008	-100	0.9		-90	Smith, John
15					-90		0	Smith, John
16							0.8	Smith, John
17	Jones, Dave	SINV00003	07/12/2008	1000	0.8		800	Jones, Dave
18					800		0	Jones, Dave
19							0.8	Jones, Dave
20		SINV00004	16/12/2008	500	0.8		400	Jones, Dave
21					400		0	Jones, Dave
22							1.1	Jones, Dave
23	Jolly, Ian	SINV00005	21/12/2008	50	1.1		55	Jolly, Ian
24					55		0	Jolly, Ian
25							0.9	Jolly, Ian

4.4.3 Product name

The first thing to notice here is that the product name has been split into two columns (A and B) because it spanned 2 columns during the data import. Therefore, before we can extract the product name, we need to join it back together in a working column. To do this we use a 'concatenation' formula. In cell I2 enter either of these formulas

=A2&B2

or =CONCATENATE(A2,B2)

Now, to extract the product name, we're going to use another fill down formula. In cell J2 enter this formula

=IF(LEFT(I2,12)="Product Name",I2,J1) and fill down

How does this formula work?

This works in a very similar manner to the previous formula, except with a different test/condition, using =LEFT(I2, 12)

Syntax =LEFT(*text*, *no of characters*)

The LEFT function will look at the leftmost characters in a text string. So here we are looking at the first 12 characters of cell I2, and seeing if they match "Product Name". If they do, then we return that product name, if not we return the value in the cell above.

The only problem we now have is that the product names all have "Product name: " at the beginning of them. We can fix this with another formula.

In the cell K2, enter the formula

=RIGHT(J2,LEN(J2)-14)

Syntax =RIGHT(*text*, *no of characters*)

The RIGHT function will extract the rightmost characters from a string. So, in this case it will extract the rightmost characters from cell J2 (which contains the product name, we just created). The problem we have is that we don't know how many characters to extract, because all products have different length names. What we want to do is extract everything but the first 14 characters (which are always "Product Name: "). Here we can use the LEN function, which simply returns the length of a text string

Syntax =LEN(*text*)

So this will take the "length -14" rightmost characters.

We don't actually need a separate column for this, we can include it in the original formula. Change the formula in J2 from

=IF(LEFT(I2,12)="Product Name",I2,J1)

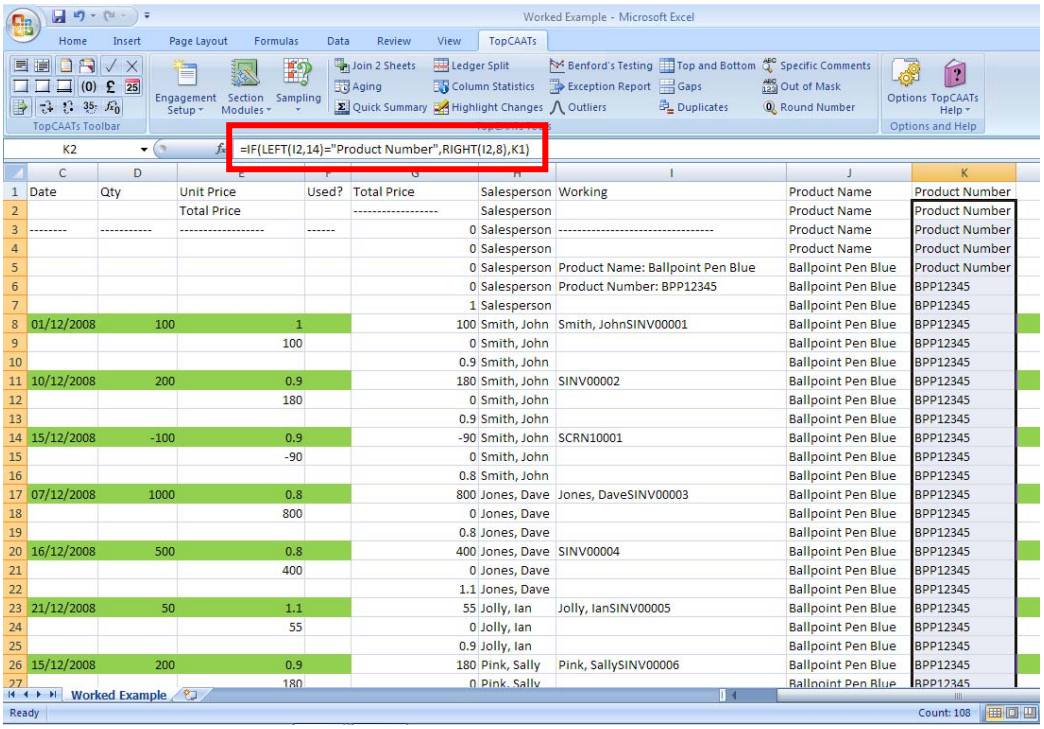
to =IF(LEFT(I2,12)="Product Name", RIGHT(I2,LEN(I2)-14),J1)

4.4.4 Product number

This works in exactly the same way as the product name, and we do not need an additional working column, as we can use the existing one (it contains both the product name and product number). The only changes are to the test (we're now looking for "Product Number") and the number of characters to extract from the right. Because the product number is a fixed number of characters we can use either of these formulas in cell K2

=IF(LEFT(I2,14)="Product Number",RIGHT(I2,LEN(I2)-16),K1)

=IF(LEFT(I2,14)="Product Number",RIGHT(I2,8),K1)



	C	D	E	F	G	H	I	J	K
1	Date	Qty	Unit Price	Used?	Total Price	Salesperson	Working	Product Name	Product Number
2			Total Price		----------	Salesperson		Product Name	Product Number
3	-------	----------	----------	------		0 Salesperson	----------	Product Name	Product Number
4						0 Salesperson		Product Name	Product Number
5						0 Salesperson	Product Name: Ballpoint Pen Blue	Ballpoint Pen Blue	Product Number
6						0 Salesperson	Product Number: BPP12345	Ballpoint Pen Blue	BPP12345
7						1 Salesperson		Ballpoint Pen Blue	BPP12345
8	01/12/2008	100	1		100	Smith, John	Smith, JohnSINV00001	Ballpoint Pen Blue	BPP12345
9			100		0	Smith, John		Ballpoint Pen Blue	BPP12345
10					0.9	Smith, John		Ballpoint Pen Blue	BPP12345
11	10/12/2008	200	0.9		180	Smith, John	SINV00002	Ballpoint Pen Blue	BPP12345
12			180		0	Smith, John		Ballpoint Pen Blue	BPP12345
13					0.9	Smith, John		Ballpoint Pen Blue	BPP12345
14	15/12/2008	-100	0.9		-90	Smith, John	SCRN10001	Ballpoint Pen Blue	BPP12345
15			-90		0	Smith, John		Ballpoint Pen Blue	BPP12345
16					0.8	Smith, John		Ballpoint Pen Blue	BPP12345
17	07/12/2008	1000	0.8		800	Jones, Dave	Jones, DaveSINV00003	Ballpoint Pen Blue	BPP12345
18			800		0	Jones, Dave		Ballpoint Pen Blue	BPP12345
19					0.8	Jones, Dave		Ballpoint Pen Blue	BPP12345
20	16/12/2008	500	0.8		400	Jones, Dave	SINV00004	Ballpoint Pen Blue	BPP12345
21			400		0	Jones, Dave		Ballpoint Pen Blue	BPP12345
22					1.1	Jones, Dave		Ballpoint Pen Blue	BPP12345
23	21/12/2008	50	1.1		55	Jolly, Ian	Jolly, IanSINV00005	Ballpoint Pen Blue	BPP12345
24			55		0	Jolly, Ian		Ballpoint Pen Blue	BPP12345
25					0.9	Jolly, Ian		Ballpoint Pen Blue	BPP12345
26	15/12/2008	200	0.9		180	Pink, Sally	Pink, SallySINV00006	Ballpoint Pen Blue	BPP12345
27			180		0	Pink, Sally		Ballpoint Pen Blue	BPP12345

4.4.5 Filter column

This is an additional column, which is going to help us apply auto filters. We're looking for an identifying feature that is common to all of the data rows, but none (or very few) of the unwanted/non data rows.

In this case we can see that the "Qty" column (column D) contains numerical values for all data rows, and contains blanks or text data for unwanted rows.

We can therefore use a very simple formula in cell L2

=ISNUMBER(D2) and fill down

This will return "TRUE" if D2 contains a number and "FALSE" if D2 doesn't contain a number

We could have used a variety of other formulas for the Filter Column, including

=IF(ISERROR(SEARCH("??.??.??",C2)),FALSE,TRUE)

This looks for a match against the date column for the format "??.??.??" (? being a single character wildcard whereas * is a multi-character wildcard) and returns "FALSE" if there is no match found or "TRUE" if a match is found.

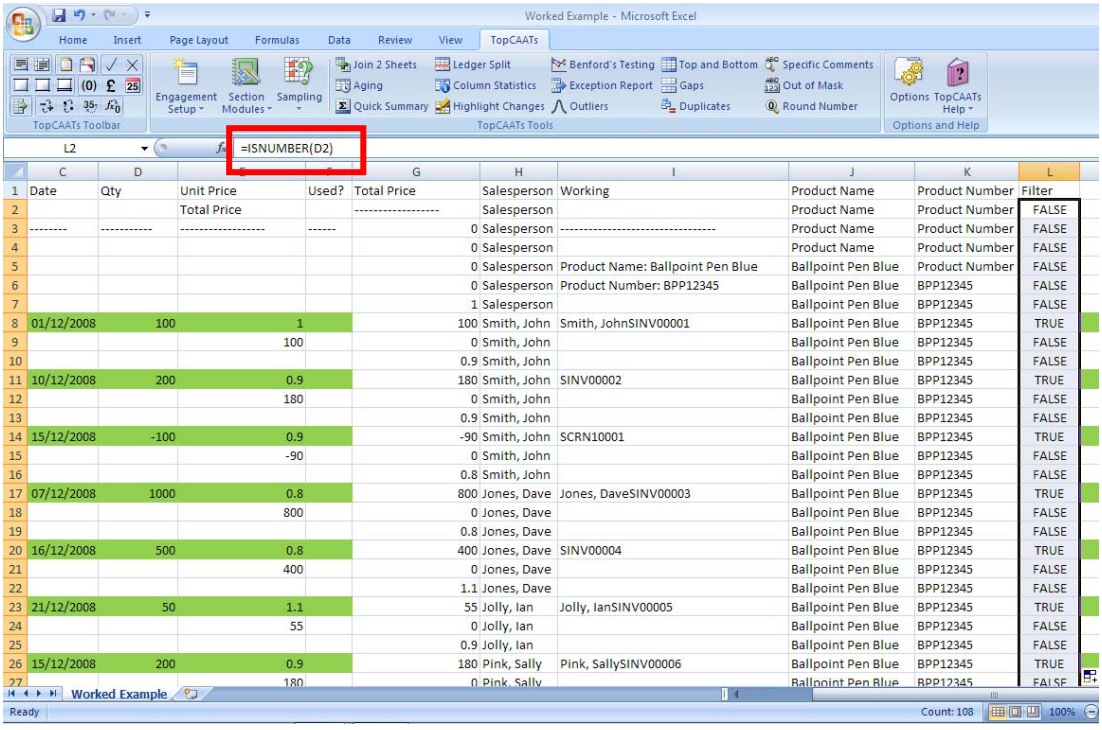
=IF(OR(LEFT(B2,4)="SINV",LEFT(B2,4)="SCRN"),TRUE, FALSE)

This looks at the left 4 characters of the invoice number for either "SINV" or "SCRN" and returns TRUE if a match is found, and FALSE if it isn't.

=ISNUMBER(C2)

This works because Excel stores dates as numbers, and hence will return TRUE if it contains a valid date and FALSE if it does not.

Please note that using any of these alternatives will alter the next step, auto filtering (you won't need to filter out the total rows, as they'll already be filtered because we are using stricter criteria in the filter column).



	C	D	E	F	G	H	I	J	K	L
1	Date	Qty	Unit Price	Used?	Total Price	Salesperson	Working	Product Name	Product Number	Filter
2			Total Price		----------	Salesperson		Product Name	Product Number	FALSE
3	--------	----------	----------	-----	0	Salesperson	----------	Product Name	Product Number	FALSE
4					0	Salesperson		Product Name	Product Number	FALSE
5					0	Salesperson	Product Name: Ballpoint Pen Blue	Ballpoint Pen Blue	Product Number	FALSE
6					0	Salesperson	Product Number: BPP12345	Ballpoint Pen Blue	BPP12345	FALSE
7					1	Salesperson		Ballpoint Pen Blue	BPP12345	FALSE
8	01/12/2008	100	1		100	Smith, John	Smith, JohnSINV00001	Ballpoint Pen Blue	BPP12345	TRUE
9			100		0	Smith, John		Ballpoint Pen Blue	BPP12345	FALSE
10					0.9	Smith, John		Ballpoint Pen Blue	BPP12345	FALSE
11	10/12/2008	200	0.9		180	Smith, John	SINV00002	Ballpoint Pen Blue	BPP12345	TRUE
12			180		0	Smith, John		Ballpoint Pen Blue	BPP12345	FALSE
13					0.9	Smith, John		Ballpoint Pen Blue	BPP12345	FALSE
14	15/12/2008	-100	0.9		-90	Smith, John	SCRN10001	Ballpoint Pen Blue	BPP12345	TRUE
15			-90		0	Smith, John		Ballpoint Pen Blue	BPP12345	FALSE
16					0.8	Smith, John		Ballpoint Pen Blue	BPP12345	FALSE
17	07/12/2008	1000	0.8		800	Jones, Dave	Jones, DaveSINV00003	Ballpoint Pen Blue	BPP12345	TRUE
18			800		0	Jones, Dave		Ballpoint Pen Blue	BPP12345	FALSE
19					0.8	Jones, Dave		Ballpoint Pen Blue	BPP12345	FALSE
20	16/12/2008	500	0.8		400	Jones, Dave	SINV00004	Ballpoint Pen Blue	BPP12345	TRUE
21			400		0	Jones, Dave		Ballpoint Pen Blue	BPP12345	FALSE
22					1.1	Jones, Dave		Ballpoint Pen Blue	BPP12345	FALSE
23	21/12/2008	50	1.1		55	Jolly, Ian	Jolly, IanSINV00005	Ballpoint Pen Blue	BPP12345	TRUE
24			55		0	Jolly, Ian		Ballpoint Pen Blue	BPP12345	FALSE
25					0.9	Jolly, Ian		Ballpoint Pen Blue	BPP12345	FALSE
26	15/12/2008	200	0.9		180	Pink, Sally	Pink, SallySINV00006	Ballpoint Pen Blue	BPP12345	TRUE
27			180		0	Pink, Sally		Ballpoint Pen Blue	BPP12345	FALSE

4.5 Applying AutoFilters

Now all we need to do is apply auto filters, so that we are left with only the data rows showing. Before we do this, it's good practice to convert all your formulas to their values, so that problems are avoided when filters are applied, or cells are extracted to a new sheet.

Select the entire worksheet, then copy (Edit > Copy) and paste values (Edit > Paste Special > Values) or use the "Convert to values" button on the TopCAATs toolbar, if you have TopCAATs installed.

Now, turn on auto filters (Data > Filter > AutoFilter), and filter out all the "FALSE" records in the filter column. If you have used the simplest filter column then you will also need to filter out the Total Rows. In Excel 2007 this can be done by unselecting the items ("Total" and "Grand Total") in column A. In previous versions of Excel you will need to apply a custom filter for "does not contain "Total"".

You should now have a clean set of data, and can copy this to a new worksheet.

If you have TopCAATs installed, simply click the "Extract visible cells" button on the TopCAATs toolbar. If you don't have TopCAATs, follow these steps

- Press Ctrl+A to select all the data
- Press Ctrl+G (or Edit > Goto) to open up the "Goto" dialog
- Click Special, select "Visible cells only" then click OK
- Press Ctrl+C (or Edit > Copy) to copy the data
- Go to a new worksheet, select cell A1, and press Ctrl+V (or Edit > Paste) to paste the data

4.6 Checking data

This is the most critical step in any data cleansing process – checking the integrity of your data, i.e. that you have correctly cleansed the data and that all records have been imported.

Total your numeric columns (e.g. units sold, total price) and ensure they agree back to the source data (which is likely to have totals printed on it, or these can be obtained from the source system).

You can also use checksums if available, e.g. totalling the invoice numbers

We cannot stress how important it is to check the data after you have cleaned it – there is no point performing data analysis on data that has been incorrectly imported or corrupted!

4.7 Deleting unwanted columns

There may be a number of unwanted columns that can be removed. In this case the first "Sales person" column is no longer needed, and nor are "Working", "Used?" and "Filter" columns, so we can delete them.

4.8 Congratulations

You've just taken a report that appeared completely unusable and cleaned it, so that you can apply data analysis tools and techniques, including using the TopCAATs tools.

Using the techniques described in this eBook you will now be able to clean almost any report. The secret is to look for logical steps that can be carried out, and then working out which formulas to use to apply these steps. This takes a bit of experience and practice, and may involve using formulas or criteria not covered in this eBook. These websites contain a vast amount of information on Excel formulae that you may find useful (the forums on MrExcel are a also great place to ask Excel related questions!):

www.ozgrid.com
www.mrexcel.com

You may also like to sign up to the MrExcel Podcast, which brings you a short (~5 min) video each day covering a different topic or challenge. You can subscribe to the podcast here, www.mrexcel.com/podcast/learnexcelpodcast.html

If you come across a report that you think looks horrible, don't be afraid! Try to break it down and look at it logically. If you were going to manually go through and pick out the real records, what would you look for? How can you translate this into logical tests you can construct in a formula?

Still stuck? We can help! We run a data cleansing service, and can clean up reports for you, and what's more we might do it for free!

We're building a video library showing how to clean a wide variety of standard system reports. If you're happy for us to publish a video showing how we have cleansed the data and we don't already have a demo for that system report, then we'll do it completely free.

We understand that your data is confidential, therefore we'll mask confidential data that appears in the video, removing anything that could identify the data source (e.g. company name, year end dates, amounts, etc). Your original data will never be published and will be kept in the strictest of confidence.

If you'd prefer us not to use your data for video demonstrations then we'll cleanse any report for a small fee. We'll also send you a video showing how we did it, so that you can follow our approach yourself in the future!

For more information on our data cleansing service please email us at cleanmydata@reinventdata.com, or visit www.topcaats.com/etc/?pid=46.

5 Formulas

This section explains the formulas used in the example in more detail

5.1 IF statements

IF allows you to test for a criteria, and then return 2 different outputs depending on the result of the test. IF is one of the most powerful and useful functions in Excel. The syntax for IF is

=IF(test, result if true, result if false)

For example, if we wanted to test whether the value in cell A2 was 100 or more, we could use this formula

=IF(A2>=100,"100 or more","less than 100")

A2>=100 is the test – i.e. is A2 greater than or equal to 100?

"100 or more" is the text to display is the outcome of the test is true (note, as this is text it is entered in double quotes)

If you need to test for more than 2 outcomes you need to "nest" IF functions. For example, if we were recalculating stock provisions based on a status code, which can either be "A" for stock that does not need a provision, "B" for 25% provision, "C" for 50% and anything else for stock that needs a 100% provision

The status code is in column C and the value is in column E

=IF(C2="A", 0, IF(C2="B", E2*0.25, IF(C2="C",E2*0.5,E2)))

This may look complicated but if we break it down it's quite simple

If C2 is A then we allocate a provision of 0. If it isn't A then we move onto the 2nd IF statement

If C2 is B then we allocate a provision of 25% (value * 0.25). If it isn't B then we move onto the 3rd IF statement

IF C2 is C then we allocate a provision of 50% (value * 0.5). If it isn't C then we allocate 100% provision (value).

5.2 AND and OR

AND and OR allow you to test multiple criteria as follows, either where all the criteria are met, or where any of the criteria are met

The syntax for these functions are as follows (n.b. you can use up to 30 criteria)

- =AND(criteria 1, criteria2…)
- =OR(criteria 1, criteria2…)

In the example spreadsheet below we have the following formulae

- C1: =AND(A1=1,B1=1)
- D1: =OR(A1=1,B1=1)

	A	B	C	D
1	1	0	FALSE	TRUE
2	1	1	TRUE	TRUE
3	0	0	FALSE	FALSE
4	0	1	FALSE	TRUE

As you can see, the formulae output TRUE or FALSE as appropriate

5.3 LEFT, RIGHT and MID

The LEFT, RIGHT and MID functions are quite similar and allow you to extract characters from the left, right or middle of a cell. The syntaxes for each are

=LEFT(cell reference, number of characters)

=MID(cell reference, start character, number of characters)

=RIGHT(cell reference, number of characters)

For example, if we wanted to extract the first 5 characters in cell A1, we could use this formula

=LEFT(A1,5)

To extract the last 5 characters from cell A1, we could use this formula

=RIGHT(A1,5)

To extract the 10 characters after the first 5 characters in cell A1, we could use this formula

=MID(A1,5,10)

To extract all characters after the first 5 characters in cell A1, you could use the following formula, assuming there are less than 999 characters in the cell

=MID(A1,5,999)

5.4 TRIM

The TRIM function removes all spaces from text, except for single spaces between words. It will remove multiple spaces, spaces at the beginning of the cell and spaces at the end of the cell. The syntax for TRIM is

=TRIM(cell reference)

For example, to apply trim to cell A1, we could use this formula

=TRIM(A1)

5.5 VALUE

The VALUE function converts text that represents a number into a number. This is most useful when your data contains 'textnumbers', which can prevent VLOOKUP functions from working. The syntax for VALUE is

=VALUE(cell reference)

For example, if cell A1 contains textnumbers, you could use this formula to convert it into real text

=VALUE(A1)

6 Techniques

This section covers the techniques we have used to clean the data, together with some additional ones you may find useful

6.1 AutoFilter

The AutoFilter tool is incredibly useful for cleansing data. AutoFilter displays only the rows that meet your specified criteria and hides rows that do not. You can also filter by more than one column. Each additional filter further reduces the data displayed.

AutoFilters can be used in two ways when cleaning data. You can either apply a filter to show the rows that you want to delete, then select those rows and delete them. Alternatively, you can apply the filter to show only the rows you want to keep, and then extract the visible cells to a new sheet.

The latter method is preferable as the source data remains untouched – if you find that you're not extracting all the data correctly, you can just go back and modify the filter. If you delete rows in error, you may lose your data and have to start again.

To apply an AutoFilter,

- Select the range of data that you wish to filter. Make sure you include all the data, including the header row in your selection.
- Apply the AutoFilter
    - Pre Excel 2007: Data > Filter > AutoFilter
    - Excel 2007: Home Tab > Sort and Filter > Filter
- Define your criteria

The AutoFilter function has been improved in Excel 2007 and makes it much easier to filter on multiple or complex criteria. However, it is possible to achieve the same results with previous versions of Excel

It is best to use AutoFilter at the end of the data cleansing process. This reduces the chances of errors occurring, as any unwanted rows still showing will stand out against the now clean data.

6.2 AutoFilter in combination with a filter column

Using the AutoFilter on its own will often enable you to filter out the rows you need, but can be made even more powerful and flexible by using it in conjunction with a filter column created using an IF statement

Say, for example, you have identified that the cells you wish to filter all have "ABC" as the 4th to 6th characters in column A, you can use the following procedure

- In a new column, enter the formula
    - =IF(MID(A1,4,3)="ABC",TRUE,FALSE)

  This will show "TRUE" if the 4th to 6th characters are "ABC" and "FALSE" if not
- You can now AutoFilter this new column to show only TRUEs

The same can be done using AND and OR functions (see above) to use more complex criteria for filtering.

6.3 Fill down formulas – dealing with subheadings and grouped data

Fill down formulas are used whenever you have information in a single cell that relates to multiple records, for example in sub-headings or when data is grouped (in our example the salespersons' names).

All fill down formulas work on the principle of testing for criteria. If that criteria is met then the new data is returned, if the criteria is not met, then the value is taken from the cell above (i.e. it returns the last cell that matched the criteria).

The test criteria will depend on your data, and can sometimes take a little imagination to determine what will work. Here are some ideas for the types of criteria you may wish to test for (assuming you're testing cell A2, and working in cell D2)

- =IF(ISNUMBER(A2),A2,D1) - *test for a numeric entry*
- =IF(LEFT(A2,3)="xyz",A2,D1) – *test the left 3 characters to see if they are "xyz"*
- Variation on above to test right or mid characters
- =IF(A2<>"",A2,D1) – *test for a non blank entry*
- =IF(ISERROR(SEARCH("xyz",A2)),D1,A2) – *test to see if A5 contains "xyz".*
    - Note that the D1 and A2 are reversed in this as the criteria will return "True" if an error is found (i.e. the search text is not found)

There may be some instances where the information required is stored below the records that it relates to. In this case you can use a "Fill up formula" which works in exactly the same way, except it returns the cell below (rather than the cell above) if the criteria is not matched.

6.4 Joining cells

There are two ways to join the text from two cells together.

The CONCATENATE function can be used, using either cell references as follows:

- =CONCATENATE(A1,B1,C1)

or with text, formulas, cell references or combinations of these, for example, if cell A1 contains “ABC12345” and cell B1 contains “67890” this formula:

- =CONCATENATE(LEFT(A1,3),”_”,B1)

will result in “ABC_67890”

The alternative method is to join bits of text using the “&” (ampersand) symbol. The above example would be written as

- =LEFT(A1,3)&”_”&B1

This can be used in many situations, but can be particularly useful if you wish to perform a VLOOKUP using data from multiple columns.

6.5 Multiple row headers / record split across 2 rows

The easiest way to deal with multiple row headers is to use a simple “=” formula, referencing the row below.

An alternative method is:

- First, copy the column containing both the first and second header data you wish to use, and paste it to a new column, giving you two columns with the same data (change the header for the new column).
- Delete the cell in row 2 of the new column by selecting the cell and using Edit > Delete (pre-2007) or Home tab > Delete (2007). When a window appear asking you whether you wish to shift cells up, shift cells left or delete the entire row or column, select ‘shift cells up’. This will now displace the column upwards by one cell, placing the second header row data on the same row as the first header row.

6.6 Convert negative to positive

Converting negative numbers to positive, and vice versa is simple, using the following formula (assuming the number you are interested in is in A1).

- =-A1

Alternatively, if you have TopCAATs installed you can use the “Reverse Polarity” button on the TopCAATs toolbar

6.7 Move "-" from right to left

Some accounting systems will place a negative sign ("-") on the right hand side of a number, which is not recognised as a number by excel. The following formula will move the negative sign from the right to the left of a number.

- =IF(RIGHT(A1,1)="-",-1*(LEFT(A1,LEN(A1)-1)))

Alternatively, if you have TopCAATs installed you can use the "Move – to front" button on the TopCAATs toolbar

6.8 Find and replace

The Find and replace function can be very useful for tidying up data.

- Press Ctrl+H (or Edit > Replace…) to launch the Find and Replace dialog
- Enter the text you want to replace
- Enter the text you want to replace it with

An example of where this is useful is deleting unwanted information in many cells. Remember in the example where we extracted the Product name and number, we used a RIGHT formula to extract the part we want?

We could have waited until the end (after we've replaced formulas with values) and used find and replace to find "Product Number: " and replace it with "" (nothing, leave the box blank), and we could have done the same with Product Name.

6.9 Deleting blank rows

The best approach with blank rows is to leave them alone until the final step and to filter them out as part of the final filtering process. However, you may wish to remove blank rows earlier on in the process, here's how

You can use AutoFilter to filter out and delete blank rows from your data.

- First, select all of the cells that contain your data and activate the AutoFilter using Data > Filter > AutoFilter (pre-2007) or Home Tab > Sort and Filter > Filter (2007)
- Apply a filter to one of the columns, using the 'Blanks' criteria
- If there are still rows showing that contain data, you will need to add another filter to one of the columns that does not contain only blanks. Continue adding filters until only blank rows are displayed.
- Select all of the rows that contain the data by clicking on the row number of the first row containing blanks (at the left of the screen) and dragging your mouse down until all rows are selected.
- Now delete these rows using Edit > Delete (pre-2007) or Home Tab > Delete (2007).

- Next, remove the AutoFilter to show all of your data using Data > Filter > AutoFilter (pre-2007) or Home Tab > Sort and Filter > Filter (2007).
- You can use AutoFilter to filter out and delete blank rows from your data.

Alternatively you can use the "Sort" feature to sort blank rows to the bottom of the data, although this should be used with extreme caution as you run the risk mixing data up and corrupting it.

6.10 Deleting subtotals / subheadings

Deleting subtotals or subheadings uses the same techniques as deleting blank rows. Again, it is best to leave these alone until the final step, and remove them as part of the final AutoFilter. In almost every case, rows that contain subtotals or subheadings will have a unique feature that differentiates them from all of the other rows. Generally they will have a unique value in one column, or will have a blank cell in a column where others do not.

- First, select all of the cells that contain your data and activate the AutoFilter using Data > Filter > AutoFilter (pre-2007) or Home Tab > Sort and Filter > Filter (2007)
- Apply a filter to one of the columns in which the subtotal / subheading row has a unique feature, for example
    - If all of the subtotal rows says "Subtotal" in column A, apply a filter to column A to show all rows where the value equals "Subtotal"
    - If all of the subheading rows has the page number (e.g. "Page 1 of 10", "Page 2 of 10" etc in column D, apply a filter to column D to show all rows where the value begins with "Page"
    - If all of the subheading rows have a blank cell in column A but text in column B, apply two filters – one to column A using the Blanks criteria, and one to column B using the Non Blanks criteria.
- If there are still rows showing that are not the subtotal / subheading, you will need to add another filter to one of the columns that contains some unique feature. Continue adding filters until only the subtotal / subheading rows are displayed.
- Select all of the rows that contain the data by clicking on the row number of the first row containing blanks (at the left of the screen) and dragging your mouse down until all rows are selected.
- Now delete these rows using Edit > Delete (pre-2007) or Home Tab > Delete (2007).
- Next, remove the AutoFilter to show all of your data using Data > Filter > AutoFilter (pre-2007) or Home Tab > Sort and Filter > Filter (2007).

We hope that you have found this eBook useful, and that it will make life easier for you (after all, at Reinvent Data our number one goal is to make your life easier!)

There are many more functions and formulas that you can use to help clean data. Excel is an extremely powerful software (and even more so when combined with TopCAATs). Once you get to grips with these techniques you'll find your productivity go through the roof!

If you find yourself stuck, then ask a colleague for help. If you're still stuck then head over to the MrExcel.com message boards and post your question there – you'll be amazed at how many helpful people there out there just waiting to answer your questions!

Keep an eye out for future eBooks, as we're planning a whole series! If you've got any suggestions for future topics please drop us an email at enquiries@reinventdata.com

And finally, if you haven't tried out TopCAATs yet, then visit www.topcaats.com and download the 30 day trial. If you use Excel on a regular basis then the chances are that TopCAATs will save you a few hours every week! You may also like to view the TopCAATs video tour at www.topcaats.com/demos/introduction.html