Activity 1 – Data quality issues

Learner guide

CONTENTS

How to Use This Workbook

Activity

Alongside this icon you'll find details of the group/individual activity or a point for everyone to discuss.

Important Idea or Concept

Generally, this icon is used to draw your attention to ideas that you need to understand by this point in the course. Let your trainer know if you do not understand or see the relevance of this idea or concept.

Key Point

This icon is used to indicate something that practitioners in this field should know. It's likely to be one of the major things to remember from the course, so check you do understand these key points.

Definition

Where a word with a very specific definition (or one that could be described as jargon) is introduced this will highlight that a definition is provided. (These words will also be found in the Glossary at the back of the workbook.)

Useful Tool

This icon indicates a technique that will help you put what you learn into practice.

Helpful Hint

This icon guides you to tips or hints that will help you avoid the standard pitfalls that await the unwary practitioner or to show you how you might increase your effectiveness or efficiency in practising what you have learnt.

Reference Material

When you have only touched briefly on a topic this icon highlights where to look for additional information on the subject. It may also be used to draw your attention to International or National Standards or Youb addresses that have interesting collections of information.

Warning

This icon is used to point out important information that may affect you and your use of the product or service in question.

Introduction

This learner guide is designed to take you through some data quality issues that you will explore and fix in Excel. You will look at each data quality dimension in detail. You will be guided with instructions on how to complete the activity in Excel.

Objectives

- Identify data quality issues in a small dataset.
- Use Excel functions to fix and handle data quality issues.

Validity

Guided activity:

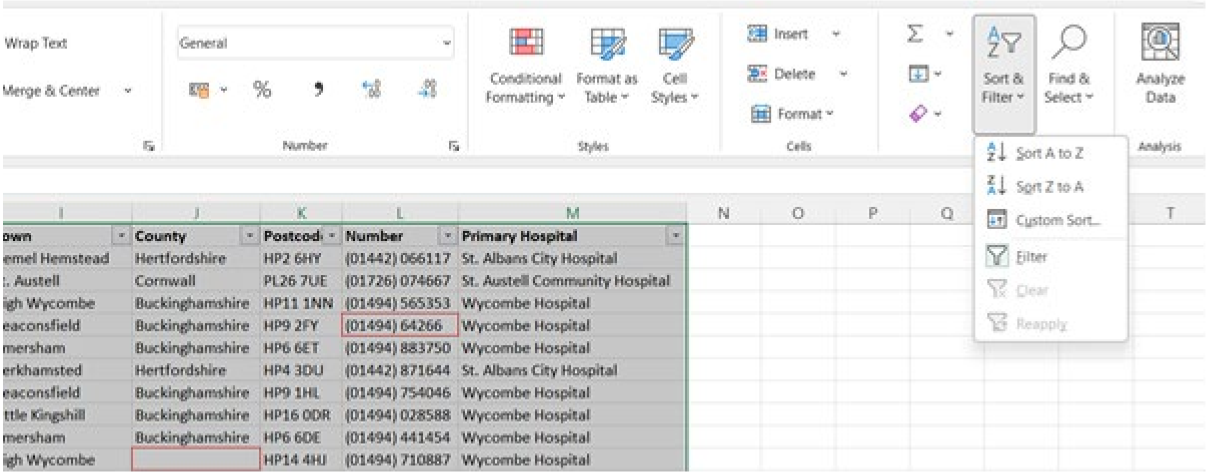
Data quality dimension: Validity

Sometimes, fields are entered manually by users, and this can lead to data not conforming to a standard value. Look at the 'Gender' column. There are two values that do not conform.

Let's find and fix the inconsistencies using the filter functionality in Excel:

1. Select the data in the whole table by clicking on any data field within the dataset and press **CTRL + A**. This will select all the data in the dataset.

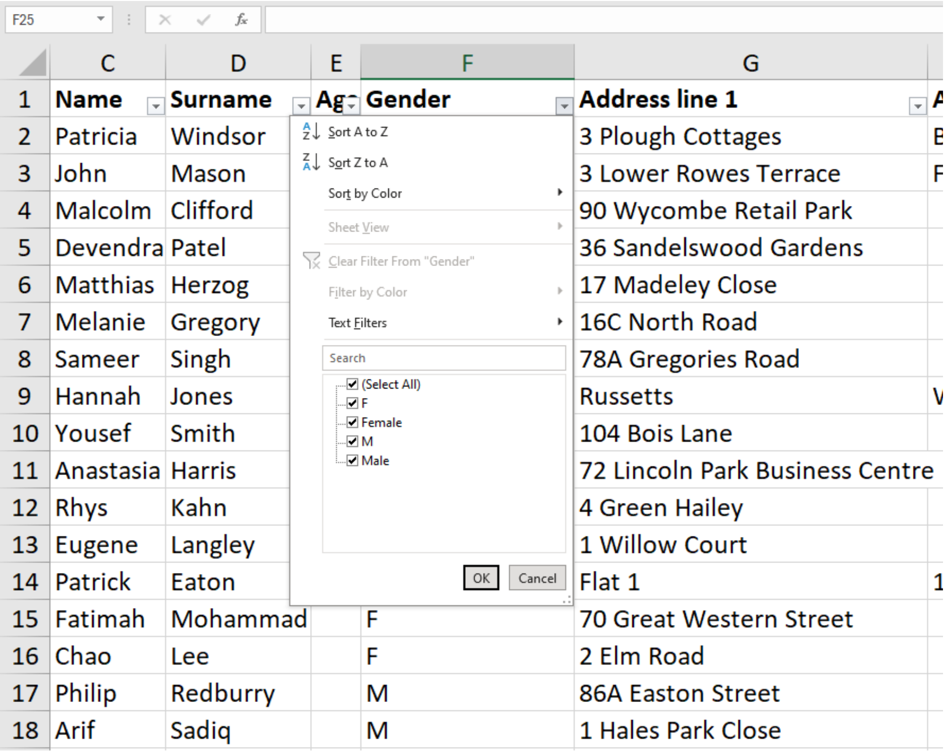
2. On the Home tab, go to **Sort & Filter > Filter**.



own	County	Postcod	Number	Primary Hospital
emel Hemstead	Hertfordshire	HP2 6HY	(01442) 066117	St. Albans City Hospital
. Austell	Cornwall	PL26 7UE	(01726) 074667	St. Austell Community Hospital
igh Wycombe	Buckinghamshire	HP11 1NN	(01494) 565353	Wycombe Hospital
eaconsfield	Buckinghamshire	HP9 2FY	(01494) 64266	Wycombe Hospital
mersham	Buckinghamshire	HP6 6ET	(01494) 883750	Wycombe Hospital
erkhamsted	Hertfordshire	HP4 3DU	(01442) 871644	St. Albans City Hospital
eaconsfield	Buckinghamshire	HP9 1HL	(01494) 754046	Wycombe Hospital
ttle Kingshill	Buckinghamshire	HP16 0DR	(01494) 028588	Wycombe Hospital
mersham	Buckinghamshire	HP6 6DE	(01494) 441454	Wycombe Hospital
igh Wycombe		HP14 4HJ	(01494) 710887	Wycombe Hospital

Figure 1: The Sort & Filter menu and dropdown on the Home tab.

3. Click the filter drop-down arrow on the 'Gender' column:



	C	D	E	F	G	
1	Name	Surname	Ag	Gender	Address line 1	A
2	Patricia	Windsor			3 Plough Cottages	B
3	John	Mason			3 Lower Rowes Terrace	F
4	Malcolm	Clifford			90 Wycombe Retail Park	
5	Devendra	Patel			36 Sandelswood Gardens	
6	Matthias	Herzog			17 Madeley Close	
7	Melanie	Gregory			16C North Road	
8	Sameer	Singh			78A Gregories Road	
9	Hannah	Jones			Russetts	V
10	Yousef	Smith			104 Bois Lane	
11	Anastasia	Harris			72 Lincoln Park Business Centre	
12	Rhys	Kahn			4 Green Hailey	
13	Eugene	Langley			1 Willow Court	
14	Patrick	Eaton			Flat 1	1
15	Fatimah	Mohammad		F	70 Great Western Street	
16	Chao	Lee		F	2 Elm Road	
17	Philip	Redburry		M	86A Easton Street	
18	Arif	Sadiq		M	1 Hales Park Close	

Figure 2: The filter drop-down open on the Gender column of the sheet.

4. A drop-down menu will appear, showing a list of all of the unique values in the column. Deselect all of the correct values (F, M), leaving all of the incorrect values selected (Female, Male). The spreadsheet will now be filtered to only show the incorrect values.

5. Fix any errors manually by typing the correct values for each one.

6. Click the column's filter drop-down arrow again and make sure all of the values listed are correct.

7. When you're done, click **Select All**, then click **OK** to show all of the rows.

There is one more issue with the dataset validity which is in the ‘Number’ column. Devendra Patel’s number is one digit short. If you add some validation to where the data is entered, perhaps this error might not occur. In this instance, it's hard to fix this data; you would have to contact the patient via post for them to contact us to rectify this issue.

Accuracy

Guided activity:

Data quality dimension: Accuracy

Let's look at the age column. This data is derived from the 'Date of Birth' data in the dataset. You can see that some of the fields are populated and some of the fields are blank. You can fix this type of data because it can be added in manually or with a formula.

If you add the data in manually, over time the data will be incorrect as time will pass but the date will not be recalculated; the data will stay static. Particularly in cases where new patients are added to the list or DOB are corrected, it would be time consuming to add it all manually.

Age = Current Date – Date of Birth (take the year value)

You can use this using the **DATEDIF** function in Excel. This calculates the number of days, months, or years betyouen two dates.

Syntax: **DATEDIF(start_date,end_date,unit)**

start_date = Date of Birth

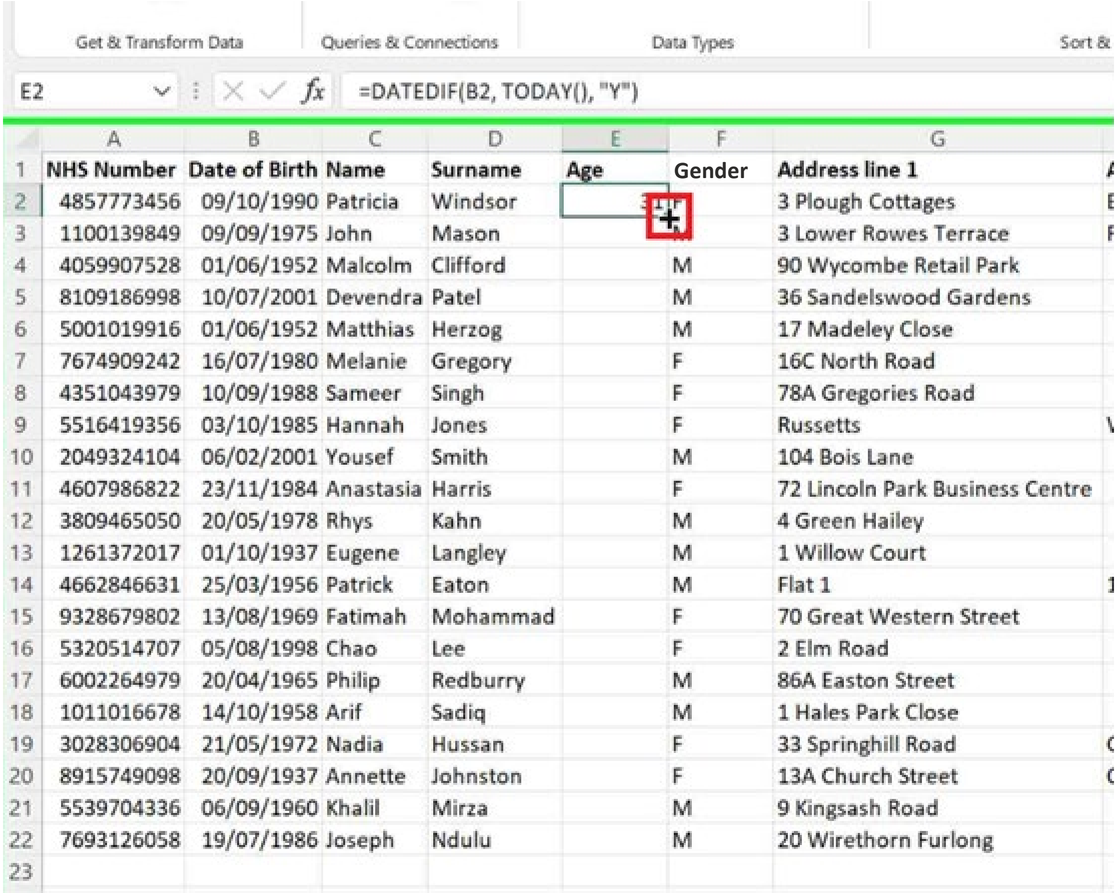
end_date = You want the **end_date** to be today's date, which you can get from the Excel function **TODAY()**

unit = **"Y"** represents the number of complete years in the period.

8. In the 'Age' column, remove any existing values in the column and in E2 use the following formula:

=DATEDIF(B2, TODAY(), "Y")

9. Use the autofill functionality in Excel to copy the formula down for the rest of the rows. If you hover over E2, the bottom right side of the cell, a plus sign (+) will appear. Double Click when the plus sign (+) appears. This should help you populate the rest of the rows within the 'Age' Column.



E2 =DATEDIF(B2, TODAY(), "Y")

	A	B	C	D	E	F	G
1	NHS Number	Date of Birth	Name	Surname	Age	Gender	Address line 1
2	4857773456	09/10/1990	Patricia	Windsor	[illegible]	F	3 Plough Cottages
3	1100139849	09/09/1975	John	Mason		[illegible]	3 Lower Rowes Terrace
4	4059907528	01/06/1952	Malcolm	Clifford		M	90 Wycombe Retail Park
5	8109186998	10/07/2001	Devendra	Patel		M	36 Sandelswood Gardens
6	5001019916	01/06/1952	Matthias	Herzog		M	17 Madeley Close
7	7674909242	16/07/1980	Melanie	Gregory		F	16C North Road
8	4351043979	10/09/1988	Sameer	Singh		F	78A Gregories Road
9	5516419356	03/10/1985	Hannah	Jones		F	Russetts
10	2049324104	06/02/2001	Yousef	Smith		M	104 Bois Lane
11	4607986822	23/11/1984	Anastasia	Harris		F	72 Lincoln Park Business Centre
12	3809465050	20/05/1978	Rhys	Kahn		M	4 Green Hailey
13	1261372017	01/10/1937	Eugene	Langley		M	1 Willow Court
14	4662846631	25/03/1956	Patrick	Eaton		M	Flat 1
15	9328679802	13/08/1969	Fatimah	Mohammad		F	70 Great Western Street
16	5320514707	05/08/1998	Chao	Lee		F	2 Elm Road
17	6002264979	20/04/1965	Philip	Redburry		M	86A Easton Street
18	1011016678	14/10/1958	Arif	Sadiq		M	1 Hales Park Close
19	3028306904	21/05/1972	Nadia	Hussan		F	33 Springhill Road
20	8915749098	20/09/1937	Annette	Johnston		F	13A Church Street
21	5539704336	06/09/1960	Khalil	Mirza		M	9 Kingsash Road
22	7693126058	19/07/1986	Joseph	Ndulu		M	20 Wirethorn Furlong
23							

Figure 3: The autofill functionality.

10. Explore the formula in E3 and E4. Notice that the reference cell changes (B3 to B4), so it calculates the age based on the corresponding row.

Completeness

Guided activity:

Data quality dimension: Completeness

Sometimes you are able to put in values by looking at other fields in the dataset (derived) – you can also find additional information online.

11. Looking at the **County** Column in the dataset, there are two missing values. Can you derive the data to populate those fields from the data you already have?

12. Looking at the **Town** column, there are two missing values. Type in **the Address line 1** and **Postcode** data into Google and see if you can find out what values can be found to populate the missing data.

Uniqueness

Guided activity:

Data Quality Dimension: Uniqueness

Looking at the dataset, you need to see if there are any duplicates. There are several ways in which you can remove duplicates from a dataset, but the one to help us look for duplicates visually is **Conditional Formatting.**

13. Select the data in **Column A**. You can click the first cell in the table column, and then press **CTRL+SHIFT+DOWN ARROW** to select fields up until where the data ends. This is particularly useful when you have 1000s of rows in your spreadsheet and you only want to select until where the data ends.

14. On the Home ribbon, find the Conditional Formatting function; it should be in the styles section.

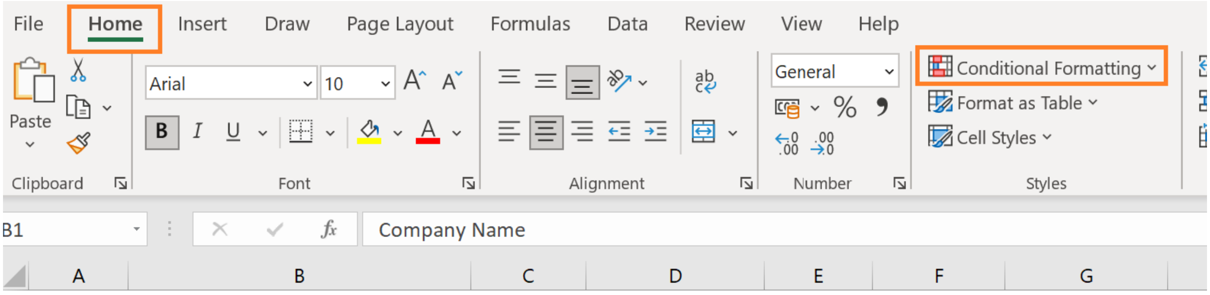


Figure 4: Conditional Formatting function.

15. Now Click the **Conditional Formatting** icon. Then go to the **Highlight Cells Rules**, then click on **Duplicate Values...**

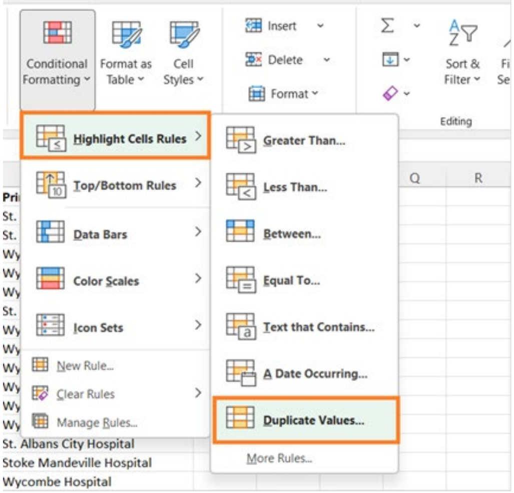
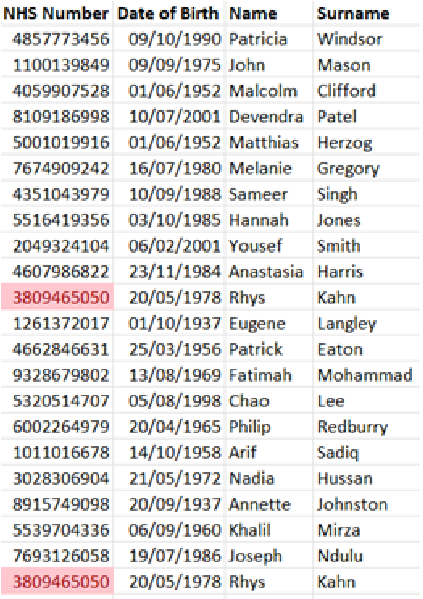


Figure 5: The Duplicate Values function under Highlight Cells Rules.

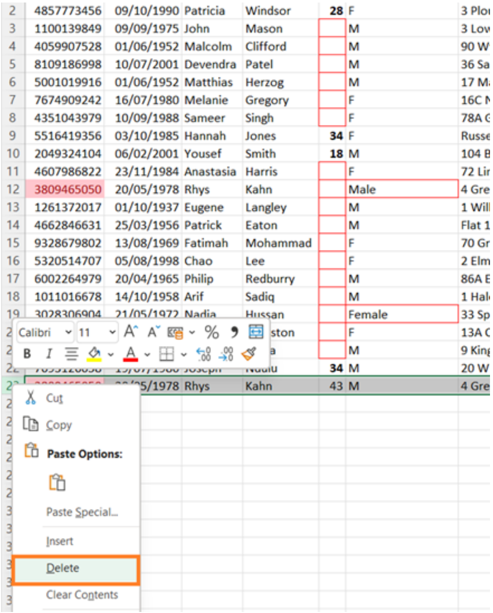
16. A pop up box should appear called **Duplicate Values**. You can change some of the parameters, but for now just click **OK**. You should now see there are two rows for Rhys Kahn:



NHS Number	Date of Birth	Name	Surname
4857773456	09/10/1990	Patricia	Windsor
1100139849	09/09/1975	John	Mason
4059907528	01/06/1952	Malcolm	Clifford
8109186998	10/07/2001	Devendra	Patel
5001019916	01/06/1952	Matthias	Herzog
7674909242	16/07/1980	Melanie	Gregory
4351043979	10/09/1988	Sameer	Singh
5516419356	03/10/1985	Hannah	Jones
2049324104	06/02/2001	Yousef	Smith
4607986822	23/11/1984	Anastasia	Harris
3809465050	20/05/1978	Rhys	Kahn
1261372017	01/10/1937	Eugene	Langley
4662846631	25/03/1956	Patrick	Eaton
9328679802	13/08/1969	Fatimah	Mohammad
5320514707	05/08/1998	Chao	Lee
6002264979	20/04/1965	Philip	Redburry
1011016678	14/10/1958	Arif	Sadiq
3028306904	21/05/1972	Nadia	Hussan
8915749098	20/09/1937	Annette	Johnston
5539704336	06/09/1960	Khalil	Mirza
7693126058	19/07/1986	Joseph	Ndulu
3809465050	20/05/1978	Rhys	Kahn

Figure 6: The two cells which show duplicate values.

17. Delete the second instance of Rhys Kahn. If you click on row number 23 on the left-hand side, next to **A23**, it will highlight the whole row. You can right click and select **Delete**:



2	4857773456	09/10/1990	Patricia	Windsor	28	F	3 Plo
3	1100139849	09/09/1975	John	Mason		M	3 Lov
4	4059907528	01/06/1952	Malcolm	Clifford		M	90 W
5	8109186998	10/07/2001	Devendra	Patel		M	36 Sa
6	5001019916	01/06/1952	Matthias	Herzog		M	17 M
7	7674909242	16/07/1980	Melanie	Gregory		F	16C N
8	4351043979	10/09/1988	Sameer	Singh		F	78A C
9	5516419356	03/10/1985	Hannah	Jones	34	F	Russe
10	2049324104	06/02/2001	Yousef	Smith	18	M	104 B
11	4607986822	23/11/1984	Anastasia	Harris		F	72 Lir
12	3809465050	20/05/1978	Rhys	Kahn		Male	4 Gre
13	1261372017	01/10/1937	Eugene	Langley		M	1 Will
14	4662846631	25/03/1956	Patrick	Eaton		M	Flat 1
15	9328679802	13/08/1969	Fatimah	Mohammad		F	70 Gr
16	5320514707	05/08/1998	Chao	Lee		F	2 Elm
17	6002264979	20/04/1965	Philip	Redburry		M	86A E
18	1011016678	14/10/1958	Arif	Sadiq		M	1 Hal
19	3028306904	21/05/1972	Nadia	Hussan		Female	33 Sp
2				ston		F	13A C
2				a		M	9 King
22					34	M	20 W
2		5/1978	Rhys	Kahn	43	M	4 Gre

Figure 7: How to use the delete function.

18. Notice how the red colour Conditional Formatting has now disappeared from your worksheet.

Consistency

Guided activity:

Data quality dimension: Consistency

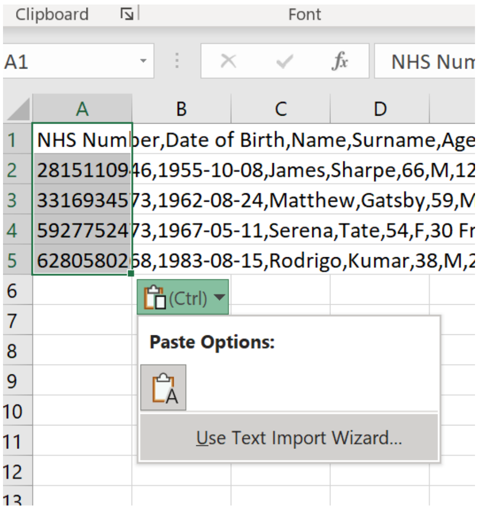
Some data has come through from another internal system. They have provided the data to us in the CSV (Comma Separated Values) format. You need to add this data to our existing spreadsheet. Luckily the columns match.

19. Select the below data in the text box, Copy (**CTRL + C**) and paste (**CTRL + V**) the data into Notepad / Notepad++ or some kind of plain text editor: The reason why you do this is to remove any formatting on the dataset before you copy the data into our spreadsheet.

```
NHS Number,Date of Birth,Name,Surname,Age,Gender,Address line 1,Address line 2,Town,County,Postcode,Number,Primary Hospital

2815110946,1955-10-08,James,Sharpe,66,M,128 Totteridge Road,,High Wycombe,Buckinghamshire,HP13 6HZ,(01494) 538080,Wycombe Hospital

3316934573,1962-08-24,Matthew,Gatsby,59,M,99 River Park Industrial Estate,,Berkhamsted,Hertfordshire,HP4 3LS,(01442) 364673,St. Albans City Hospital

5927752473,1967-05-11,Serena,Tate,54,F,30 Frogmore Street,,Tring,Hertfordshire,HP23 5AZ,(01296) 385744,St. Albans City Hospital

6280580268,1983-08-15,Rodrigo,Kumar,38,M,211 Dashwood Avenue,,High Wycombe,Buckinghamshire,HP12 3DB,(01494) 134402,Wycombe Hospital
```

20. Copy (**CTRL + C**) the data from your Notepad/test editor and paste (**CTRL + V**) the data into your NHS dataset spreadsheet, in the next empty row. For example, in column A row 23.

21. Click on the **(Ctrl)** box to expand the options. You should see **Use Text Import Wizard...** Click on this option.



A1 · NHS Num

```
1 NHS Number,Date of Birth,Name,Surname,Age
2 2815110946,1955-10-08,James,Sharpe,66,M,12
3 3316934573,1962-08-24,Matthew,Gatsby,59,M
4 5927752473,1967-05-11,Serena,Tate,54,F,30 Fr
5 6280580268,1983-08-15,Rodrigo,Kumar,38,M,2
```

Figure 8: The Ctrl box pop-up box that shows the Paste options and Use Text Import Wizard options.

(If you are having trouble getting this option, go to the data tab and click on the Text to Columns button).

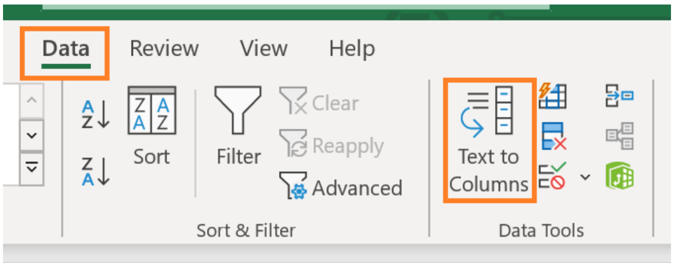
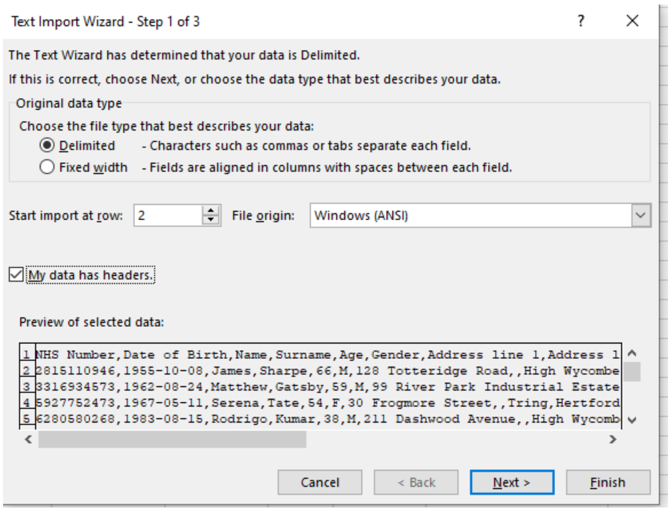


Figure 9: The Text to Columns option under the Data tab.

22. The Text Import Wizard examines the text that you are importing and helps ensure that the data is imported in the way that you want. The items in the data pasted is separated by commas, there it is delimited; you need to select the **Delimited** option. If you have also included the headings, **start import at row** 2, else leave it at 1. Tick **My data has headers** (if available) and click **Next**.



Text Import Wizard - Step 1 of 3

The Text Wizard has determined that your data is Delimited.

If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

Choose the file type that best describes your data:

● Delimited - Characters such as commas or tabs separate each field.

○ Fixed width - Fields are aligned in columns with spaces between each field.

Start import at row: 2 File origin: Windows (ANSI)

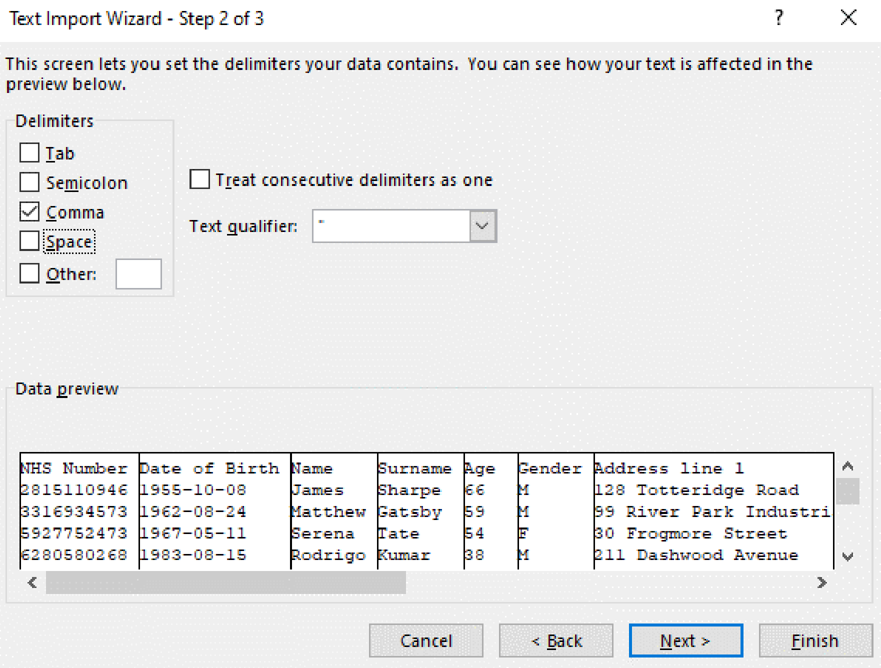
☑ My data has headers.

Preview of selected data:

```
1 NHS Number,Date of Birth,Name,Surname,Age,Gender,Address line 1,Address l
2 2815110946,1955-10-08,James,Sharpe,66,M,128 Totteridge Road,,High Wycombe
3 3316934573,1962-08-24,Matthew,Gatsby,59,M,99 River Park Industrial Estate
4 5927752473,1967-05-11,Serena,Tate,54,F,30 Frogmore Street,,Tring,Hertford
5 6280580268,1983-08-15,Rodrigo,Kumar,38,M,211 Dashwood Avenue,,High Wycomb
```

Figure 10: The Text Import Wizard - Step 1 of 3.

23. You need to let the import wizard know that the delimiter for our dataset is only commas, so be sure to only tick this option. If any other options are ticked, you need to untick them. Once done, click **Next**.



Text Import Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

- □ Tab
- □ Semicolon
- ☑ Comma
- □ Space
- □ Other:

□ Treat consecutive delimiters as one

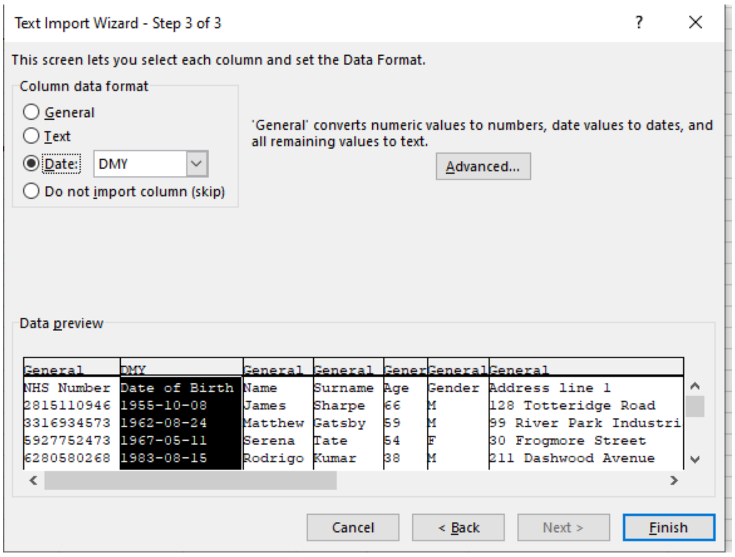
Text qualifier: "

Data preview

NHS Number	Date of Birth	Name	Surname	Age	Gender	Address line 1
2815110946	1955-10-08	James	Sharpe	66	M	128 Totteridge Road
3316934573	1962-08-24	Matthew	Gatsby	59	M	99 River Park Industri
5927752473	1967-05-11	Serena	Tate	54	F	30 Frogmore Street
6280580268	1983-08-15	Rodrigo	Kumar	38	M	211 Dashwood Avenue

Figure 11: The Text Import Wizard - Step 2 of 3.

24. If you look at the dataset, you can see the 'Date of Birth' column contains a different format than our data. Our data come to us as DD/MM/YYY, but this data is YYYY-MM-DD. Select the 'Date of Birth' Column and choose the **Column data format** to be Date **DMY**. Now click **Finish**.



Text Import Wizard - Step 3 of 3

This screen lets you select each column and set the Data Format.

Column data format

- ○ General
- ○ Text
- ● Date: DMY
- ○ Do not import column (skip)

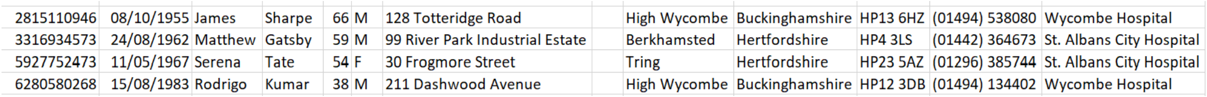
'General' converts numeric values to numbers, date values to dates, and all remaining values to text.

Data preview

General	DMY	General	General	Gener	General	General
NHS Number	Date of Birth	Name	Surname	Age	Gender	Address line 1
2815110946	1955-10-08	James	Sharpe	66	M	128 Totteridge Road
3316934573	1962-08-24	Matthew	Gatsby	59	M	99 River Park Industri
5927752473	1967-05-11	Serena	Tate	54	F	30 Frogmore Street
6280580268	1983-08-15	Rodrigo	Kumar	38	M	211 Dashwood Avenue

Figure 12: The Text Import Wizard - Step 3 of 3.

25. Your data should look like this:



2815110946	08/10/1955	James	Sharpe	66	M	128 Totteridge Road		High Wycombe	Buckinghamshire	HP13 6HZ	(01494) 538080	Wycombe Hospital
3316934573	24/08/1962	Matthew	Gatsby	59	M	99 River Park Industrial Estate		Berkhamsted	Hertfordshire	HP4 3LS	(01442) 364673	St. Albans City Hospital
5927752473	11/05/1967	Serena	Tate	54	F	30 Frogmore Street		Tring	Hertfordshire	HP23 5AZ	(01296) 385744	St. Albans City Hospital
6280580268	15/08/1983	Rodrigo	Kumar	38	M	211 Dashwood Avenue		High Wycombe	Buckinghamshire	HP12 3DB	(01494) 134402	Wycombe Hospital

Figure 13: How the data should look in the spreadsheet.

If your headers got imported, just delete that line.

26. There is only one problem you have to fix, which is the ‘Age’ column. The data is static, so the data won’t change when the dates change. You need to select E22, this should have the formula you put into the dataset. If you hover at the bottom of E22 on the right-hand side, the plus arrow should come up and you can double click for the formula to replace the data in the ‘Age’ column.