Applying Behavioural Insights to the Early Years: A Toolkit



**Data cleaning and merging guide**

### **Introduction**

Spreadsheets often contain data that is readable to humans but that Excel cannot interpret. Data cleaning, which formats data for Excel to interpret, is essential for data analysis.

This guide walks you through how to clean and merge your data in Excel. Use it with the Excel version of the guide. You can also find a short video walkthrough guide on the toolkit website.

The example we'll use here is a fictional study exploring whether changing the wording of a letter can increase the number of parents applying for the 2 year old free childcare offer. The steps would be similar when you run your own study.

**Why do we need to merge data?**

Your outcomes data may be held separately from your randomised data (whether the parents are in the intervention or control group). For analysis, we need to merge them. This means ensuring that each person's intervention assignment is matched to their outcome. You may do something similar to this if you match the DWP list with your internal datasets to send your communications out.

**Note**: this is often not as simple as combining the columns side by side into a single spreadsheet, because the order of people in each dataset may be very different, leading to mismatched rows.

Before you get started, which variable will be used for merging. This variable should:

1. Appear in both the randomisation and outcome datasets
2. Be unique for each person in the data, such as their full name or a unique ID number

As an example, we will use each person’s full name in this guide.

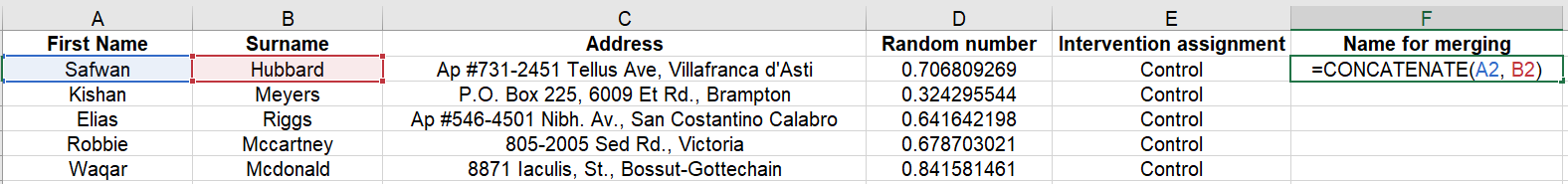
### **Step one:** Creating a merging variable

Open the Excel data cleaning guide and click on the ‘Randomised data’ tab. You’ll see a list of 100 fake parents and whether they are in the control or intervention group.

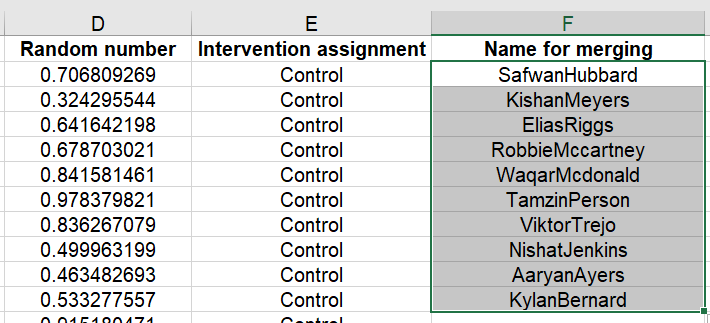
To create the merging variable, we need to join parents’ first and last names together.

Type the following command into the first cell of the ‘Name for merging’ column. This joins cells A2 and B2 together.

=CONCATENATE(A2, B2)



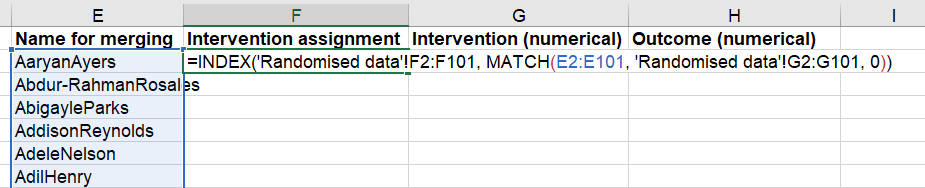
Next, hover your cursor on the bottom right of the cell until a small black cross appears. Click and drag the cross down to the bottom of the column to do the same for every parent in the list.



### **Step two:** Performing the merging

Go to the ‘Outcome data’ sheet in the Excel guide. In the ‘Intervention assignment’ column, enter the following command in the first blank cell after the title:

=INDEX('Randomised data'!F2:F101, MATCH(E2:E101, 'Randomised data'!G2:G101, 0))



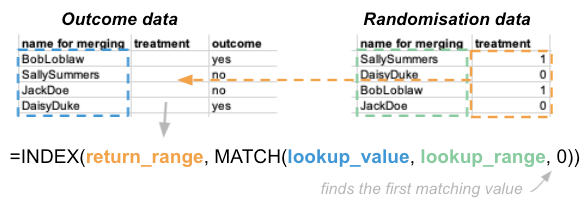
Once you’ve entered this command, the data should appear in the whole of the column. If it doesn’t, select the bottom right hand corner of the row where you entered the command and drag the fill handle Fill handle down to the final row.

When you do this for your own dataset, the command will be slightly different. The general form the command needs to take is:

=INDEX(return\_range, MATCH(lookup\_value, lookup\_range, 0))

The key parts of this command are:

* The **return\_range** is the column of data that you want included in the outcome dataset (i.e. whether parents are in the intervention group or control group);
* The **lookup\_value** is the merging variable in the outcome dataset (e.g. the cleaned full name data).
* The **lookup\_range** is the merging variable in the randomisation dataset.
* The zero (0) argument tells Excel to find the first matching value.



Your datasets might be held in different Excel files, or in different sheets of the same Excel file. To access data from different files or sheets, use single quotation marks and the exclamation mark symbol:

* **For a different excel file:** *[Workbook\_name]Sheet\_name!Cell\_location (e.g. [outcomes\_data.xlsx]2 (example) randomisation data!A2:A10)*
* **For a different sheet in the same file:** *='Name\_of\_sheet'!Cell\_location* *(e.g. ‘2 (example) randomisation data’!A2:A10)*

Finish this step by checking that everything worked as expected (e.g. by doing a few spot checks of both the randomised and outcomes data and ensuring they have merged properly).

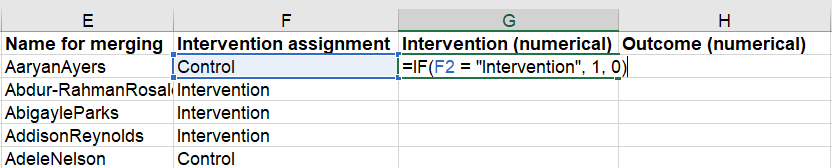
### **Step three:** Re-coding your data

Re-coding variables means changing the way that the information in the column is recorded. Specifically, we need to change text into 0’s and 1’s so that Excel can use them for analysis.

In this example, our parents are either in the control group or the intervention group. We need to change this text so that it is formatted as a binary variable with only 2 possible values (0 = control group, 1 = intervention group).

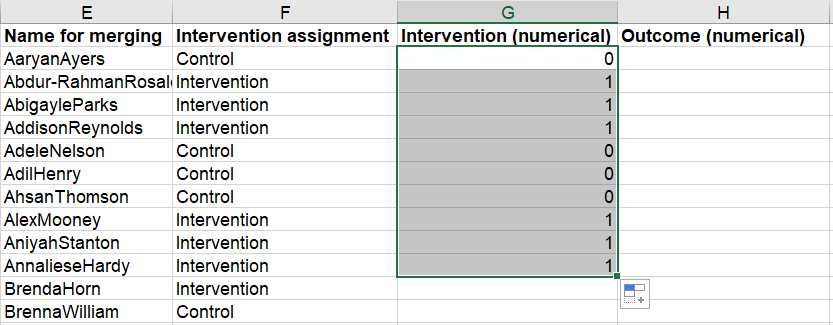
To practice this, go to the ‘Outcome data’ sheet in the Excel guide. Type the following command into the first row of the ‘Intervention (numerical)’ column:

=IF(F2 = "Intervention", 1, 0)



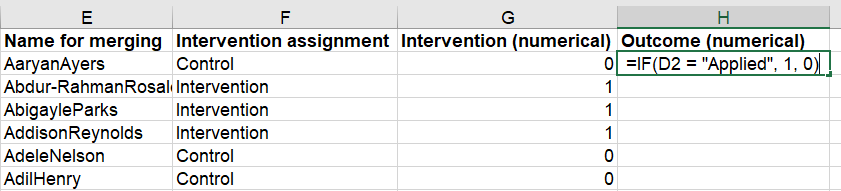
This command tells Excel to enter a 1 if the word “Intervention” appears in cell F2. Otherwise, it will enter a 0. Because this participant is in the control group, a 0 will appear.

Repeat this process for every participant in the list. Select the bottom right hand corner of the cell and drag down to the bottom. A series of 0s and 1s will appear, corresponding to which group parents are in.



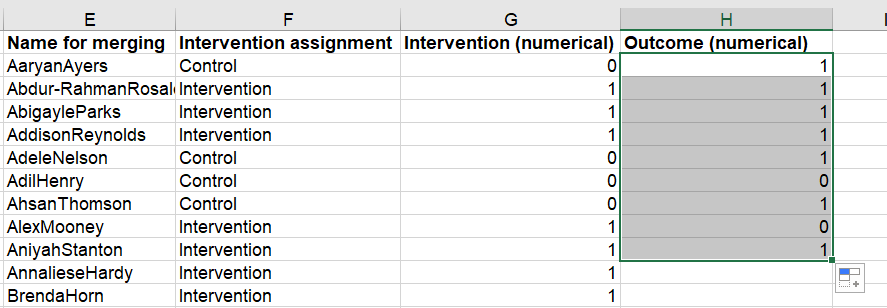
Finally, we’ll do the same for the outcomes data. In the first cell of the ‘Outcome (numerical)’ column, type the following command:

=IF(D2 = "Applied", 1, 0)



**Note**: In our example the outcome data has 3 possible values: “Applied”, “Declined”, “Did not reply”. This command will convert the data to 2 values, where 1 = “Applied” and 0 = both “Declined” and “Did not reply”.

Repeat this process for every participant in the list by selecting the bottom right hand corner of the cell and dragging down to the bottom (just as you did for the intervention assignment column).



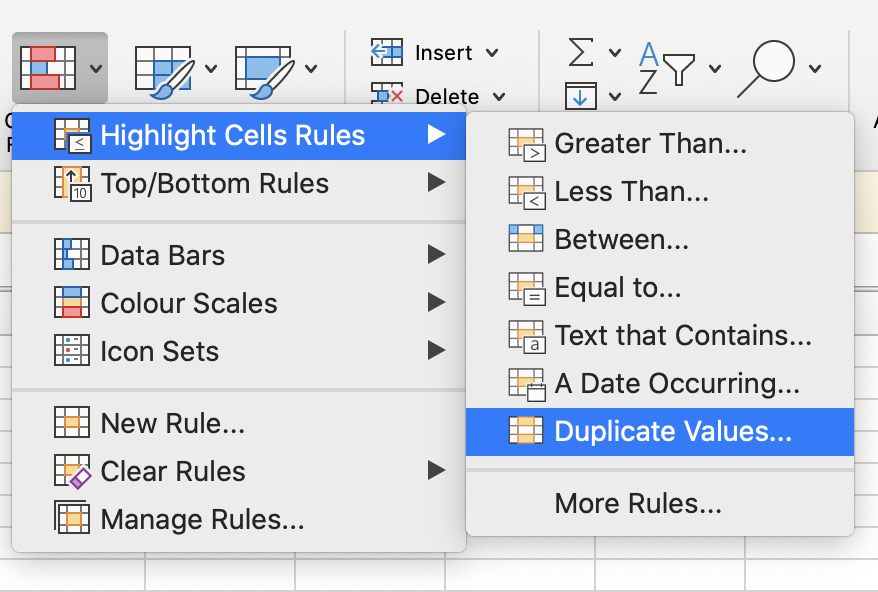
That’s it - you’ve merged and cleaned this list. You can now follow the same steps to do the same for your own dataset. Once that’s done, you can now move onto analysis (details in our separate guide).

### **Other data cleaning tips**

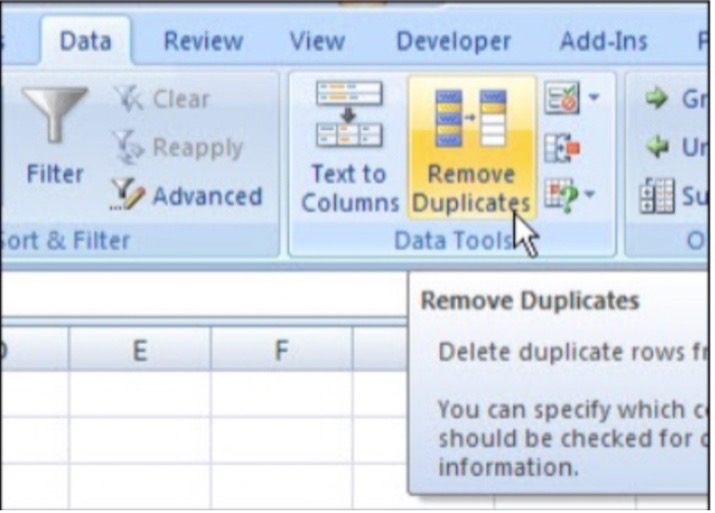
Different datasets store information in different ways, and there are lots of separate steps you might need to take to clean yours in practice. Here are some things you can do:

* **Look for and remove duplicates:** Check names and addresses to see if there are duplicates. If there are completely identical duplicate rows, delete them. There will be lots of duplicate first names and last names, so make sure the full name is a duplicate before removing the row.

*Use the* ***“Highlight Duplicates”*** *option to find duplicate cells in a column or table.*



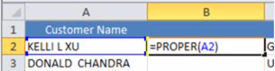
*Use the* ***“Remove Duplicates'' option*** *to automatically delete duplicate values in a column or table (You might need to delete them manually, e.g. if you want to keep the earlier entry).*

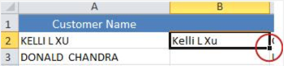


*Use* ***“Find & Replace”*** *to identify specific characters and replace them with other characters.*

* **Capitalisation:** Standardise use of capitalisation (especially for names)

*Change the appearance of cell values using UPPER(), LOWER(), and numeric formatting. For names, you will want to use “proper” capitalisation (use the function “PROPER()”).*





* **Removing other characters:** Remove extra spaces & non-print characters
* Use the “TRIM()” function to remove spaces from text.
* Use the “LEN()” function to identify the number of characters in any given cell. You can use this to identify names or addresses that are way too long or too short.

| **Some other Excel functions that may be useful:**   * CONCATENATE(string\_1, string\_2,...etc.) – joins multiple strings from different cells together (e.g. Creating “Full Name” out of first name and last name). * RIGHT(string, number\_of\_characters) – returns a desired number of characters from the right-hand side of a string. You can choose how many characters to return. For example, this can be useful if from “Mr. Bob”, you want to only get “Bob” and drop the title: =RIGHT("Mr. Bob",3). * LEFT(string, number\_of\_characters) – as above, but from the left-hand side of a string. * MID(string, number\_of\_characters) – returns characters from the middle of a string. |
| --- |