Excel: Clean Up Messy Data

Carolyn Pecoskie, Metadata & Electronic Resources Librarian, McGill University
carolyn.pecoskie@mcgill.ca

This workshop is being offered on August 9, 2022 from 3PM-4PM as part of the first BiblioTECH Jumpstart Program.

Table of Contents

Attributions

Parts of this content are taken from these Digital Scholarship workshops by Alisa Rod and Clara Turp: https://github.com/Digital-Scholarship-Hub/WorkingWithDataInExcel and https://github.com/Digital-Scholarship-Hub/Excel_Intermediate

Themes of this workshop are mapped to components of this Library Carpentry OpenRefine workshop: https://librarycarpentry.org/lc-open-refine/ (sections 1, 2, 4, 5, 7, 8). This workshop can be considered a companion to the OpenRefine: Clean Up Messy Data workshop by Robin Desmeules.

Thank you to Nadine Desrochers for her assistance with the French translation of this workshop And thank you to the first cohort of the BiblioTECH Jumpstart Program for their contributions and suggestions.

Further Reading:
For more advice for working with spreadsheets as an information professional, please see: "Tidy Data for Librarians" by The Carpentries: https://librarycarpentry.org/lc-spreadsheets/

Learning Goals

The goal of today's workshop is to introduce you to useful tips, tricks, and tools in Excel, to begin to guide your work with Excel as an information professional. This session will be by no means comprehensive, in terms of covering everything that you may ever need to use in Excel, but the hope is that by the end of the session you will:

- Be aware of a range of tools and functions that (in the experience of the presenter) are very useful for library work
- Be aware of some helpful tips and tricks to save time and make the most of what Excel can do for you
- Feel more confident in your ability to navigate within Excel, and to look to Google, the Microsoft Excel help site, and other sources whenever you need to find a new tool or function or troubleshoot an error

This workshop is being offered just prior to the OpenRefine: Cleaning Up Messy Data workshop. We are offering these workshops in conjunction because Excel and OpenRefine can both be very helpful for cleaning up and working with datasets, but they offer quite different tools and functions. We hope that by the end of today, you will have a sense of when to use either tool, based on what you would like to do with your data.

Setup and Introduction to Excel

Excel is one of the most commonly used proprietary spreadsheet software, and is part of the Microsoft Office suite of software. LibreOffice, OpenOffice.org, and Gnumeric are other examples of open source spreadsheet programs. They have similar functionalities, although some might be represented slightly differently.

You should have access to Excel through your university:

- McGill
- UdeM

If you don’t have access to it, please reach out to someone on the BiblioTECH organizing committee.

For the workshop, you will need to download a dataset. Click here to download the dataset.

Quick Tips and Tricks

File Types

Microsoft help page: File formats that are supported in Excel

You can work with a variety of file types in Excel, including:

- Excel file formats (.xlsx or the legacy .xls are the most popular)
- Text file formats (such as .txt or .csv)

Excel will encourage you to save files in one of their own formats, but sometimes it is better to keep it in the original format (such as .txt, .csv, etc.), especially if you want certain pieces of data to stay formatted in a very particular way (especially dates, as Excel can tend to get a bit overzealous in trying to reformat your dates for you).

You can save an Excel file in another format by clicking **File** > **Save As** and choosing one of the available file formats from the drop-down menu.

From Microsoft: “To open a file that was created in another file format, either in an earlier version of Excel or another program, click **File** > **Open**. If you open an Excel 97-2003 workbook, it automatically opens in **Compatibility Mode**. [...] However, you also have the option to continue to work in **Compatibility Mode**, which retains the original file format for backwards compatibility.”

Data Types

Microsoft help page: Available number formats in Excel

In Excel, you can change the “format” of a cell, so that it will recognize its contents as things like currency, percentages, decimals, dates, text, etc.

To do this, select a cell or cell range that you wish to change. Right-click on the selection and choose **Format Cells…** Within the **Format Cells** menu, choose the **Number** tab (this is the first tab so likely to appear by default). You can choose which number format you wish to use, and for certain number formats, you can customize the format even further. From the Microsoft help page linked above, here are the available number formats in Excel:

Format	Description
General	The default number format that Excel applies when you type a number. For the most part, numbers that are formatted with the **General** format are displayed just the way you type them. However, if the cell is not wide enough to show the entire number, the **General** format rounds the numbers with decimals. The **General** number format also uses scientific (exponential) notation for large numbers (12 or more digits).
Number	Used for the general display of numbers. You can specify the number of decimal places that you want to use, whether you want to use a thousands separator, and how you want to display negative numbers.
Currency	Used for general monetary values and displays the default currency symbol with numbers. You can specify the number of decimal places that you want to use, whether you want to use a thousands separator, and how you want to display negative numbers.
Accounting	Also used for monetary values, but it aligns the currency symbols and decimal points of numbers in a column.
Date	Displays date and time serial numbers as date values, according to the type and locale (location) that you specify. Date formats that begin with an asterisk (*) respond to changes in regional date and time settings that are specified in Control Panel. Formats without an asterisk are not affected by Control Panel settings.
Time	Displays date and time serial numbers as time values, according to the type and locale (location) that you specify. Time formats that begin with an asterisk (*) respond to changes in regional date and time settings that are specified in Control Panel. Formats without an asterisk are not affected by Control Panel settings.
Percentage	Multiplies the cell value by 100 and displays the result with a percent (%) symbol. You can specify the number of decimal places that you

Format	Description
	want to use.
Fraction	Displays a number as a fraction, according to the type of fraction that you specify.
Scientific	Displays a number in exponential notation, replacing part of the number with E+n, where E (which stands for Exponent) multiplies the preceding number by 10 to the nth power. For example, a 2-decimal **Scientific** format displays 12345678901 as 1.23E+10, which is 1.23 times 10 to the 10th power. You can specify the number of decimal places that you want to use.
Text	Treats the content of a cell as text and displays the content exactly as you type it, even when you type numbers.
Special	Displays a number as a postal code (ZIP Code), phone number, or Social Security number.
Custom	Allows you to modify a copy of an existing number format code. Use this format to create a custom number format that is added to the list of number format codes. You can add between 200 and 250 custom number formats, depending on the language version of Excel that is installed on your computer. For more information about custom formats, see Create or delete a custom number format.

Autofill

Microsoft help page: Fill data automatically in worksheet cells

The **Autofill** feature is very useful in Excel, for filling cells with data that follow a pattern.

1. First, select about three cells that are next to each other to use as a basis for filling additional cells that will be next to these original three in the spreadsheet. Patterns could include things like “1, 2, 3”, “15, 30, 45”, or a series of formulas.
2. Drag the fill handle:
3. If needed, click **Auto Fill Options** for additional options.

Note that when you go to autofill a formula, you will need to add the appropriate notation (“$”) in the case where you want to have an absolute reference, in terms of either column or row or both. For more information, please see the Microsoft help page: Switch between relative and absolute references

Freeze Panes

Microsoft help page: Freeze panes to lock rows and columns

Sometimes when you have complex data it is really useful to have the headers stay at the top of the page, even if you scroll through rows of data. Excel has a functionality that allows you to do this really quickly. This functionality is called **Freeze Panes**.

You want to click on the **View** tab, and go to the **Freeze Panes** sub-section. You can then choose if you want to freeze the top row, the first column, or any two panes of your choice.

Insert or Delete Rows and Columns

Microsoft help page: Insert or delete rows and columns and Insert one or more rows, columns, or cells in Excel for Mac

The easiest way to insert or delete an entire row or column is to right-click on the top of the row or column (i.e. click on the letter or number that denotes that row or column) and choose to **Insert** or **Delete**. New rows will be inserted *above* the selected row, and new columns are inserted to the *left* of the selected column. Cell references should automatically adjust based on where cells are shifted due to an insert or delete.

If you wish to insert or delete multiple rows or columns at once, this is also possible. When inserting, you simply need to select the same number of rows or columns as you want to insert. For example, to insert five blank columns (or rows), you need to select five columns (or rows). It is okay if the selections contain data, because the blanks will be inserted to the left (for columns) or above (for rows) compared to the selection.

Troubleshooting Errors

Excel always gives you some information when you make an error. The first thing to notice is that error messages always start with a hashtag (#) and finish with a punctuation sign (often ! or ?). “**#DIV/0!**” is the message that appears when you try dividing a number by zero. This is because empty cells are treated as zeros for the purposes of mathematical calculations. The error message is short but gives you some information. Other common error messages are: **#NAME?**, **#NULL!**, **#REF!**, **#VALUE!**. If you see those, don’t panic, and try copying the error in Google. You can also hover your cursor over the small green triangle that appears in the top left corner of the cell with an error message. This will reveal a yellow caution sign with an explanation point, which is a drop-down menu that you can click to gain further information about the error or help.

Sorting and Filtering

Sorting and **Filtering** are related functions, but they have different uses. The main difference between the two is that **Sorting** will rearrange your data, ultimately changing the order in which the data is appearing in your spreadsheet. **Filtering**, on the other hand, temporarily changes the view of the data on the sheet, but can easily be toggled on or off, thereby returning the spreadsheet to its original state and original order very easily. We'll take a look at how to use each of these functions in more detail.

Sorting

Microsoft help page: Sort data in a range or table

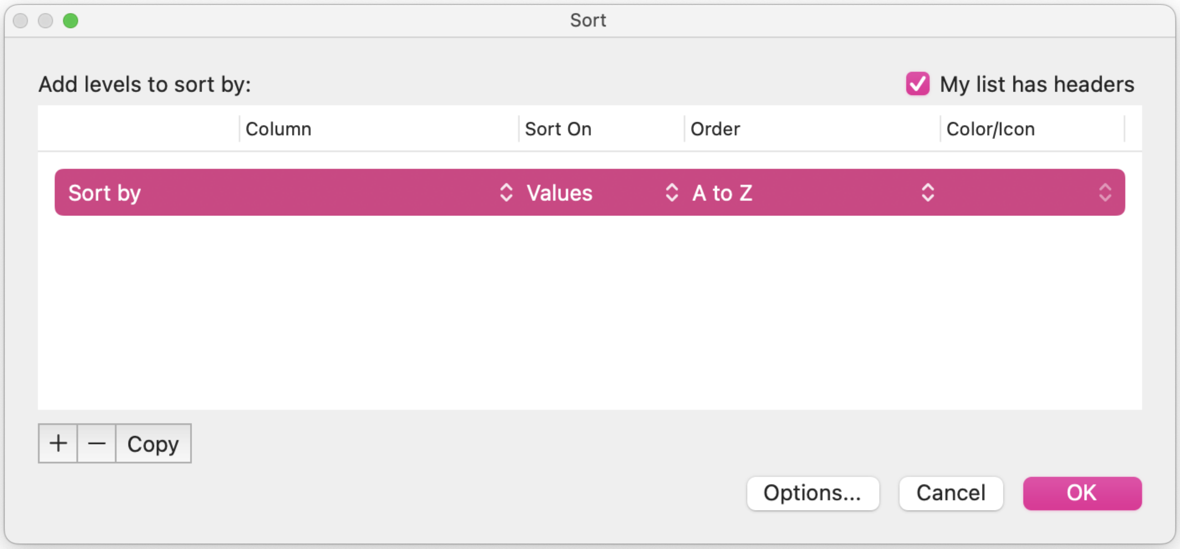
There are many reasons why you may want to sort your data. You may want a list of alphabetized names, or you may want observations organized by date so you can enter more variables.

Currently, the dataset is sorted by “interview_date”. We are going to sort this data set by “village” instead. To do this, we need to highlight all of the data including the variable names (in the top row). Any cells that are not selected will remain in the exact order they are now. There are many ways to highlight the data. We will click on cell A1. Then we will press **CTRL (or Command on a Mac) + Shift + (the right arrow)**. This will highlight the cells from A1 to BJ1.

Now we want to highlight all of the rows below those cells. Now press **CTRL + Shift + (the down arrow)**. Now all of the data should be selected.

Now we want to sort the data. In the **Home** tab, click on the arrow next to the **Sort & Filter** button. Then choose the **Custom Sort** option.

You should see the following pop-up window:



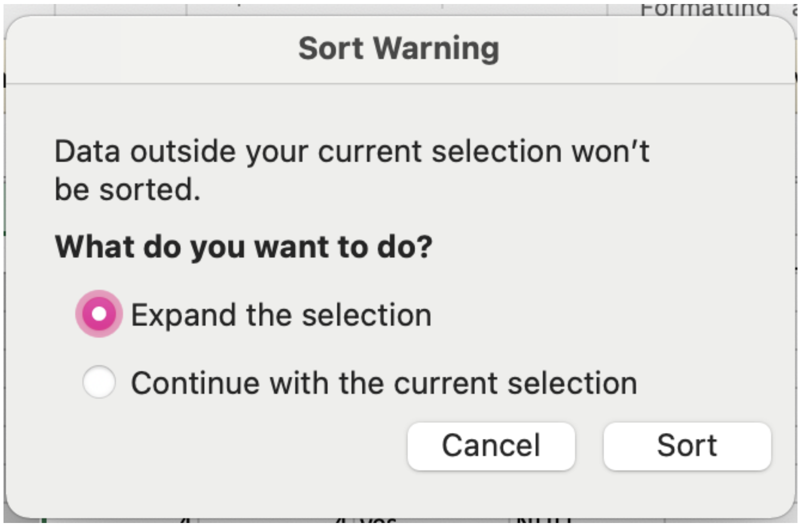
The first thing you want to do is check **My list has headers** so your variable names are not sorted like an observation. When alphabetizing the village names, you do not want the variable name “village” to appear among the list of names.

Then next to **Sort by** you want to choose “village” under the **Column** drop down menu. If you wanted to sort by multiple variables, you could click the plus sign to add a second variable to sort by. In our case, we will stick with one, and then press **OK**.

Now you should notice that the dataset has been sorted.

Sort Warning

Another way to sort by a single variable is to just select the single column you want sorted, and then on the **Home** tab, click the down arrow next to **Sort & Filter**. In this case, you will be presented with a warning:



Sort Warning

Data outside your current selection won't be sorted.

What do you want to do?

● Expand the selection

○ Continue with the current selection

Expand the Selection: 9 times out of 10 this is what you want to do, otherwise the column you have selected will be the only one that gets rearranged, and (potentially critical) relationships between the columns will be lost.

Filtering

Microsoft help page: Filter data in a range or table and Filter for unique values or remove duplicate values

Filtering is about temporarily changing your view of the spreadsheet, but not making a permanent change to the order of any content. You can toggle the **Filtering** function on or off for single or multiple columns at a time. To do this, select the column(s) that you want to apply filtering to. You can hold down the **CTRL (or Command key on Mac)** to select multiple columns that are not necessarily side by side. Then, on the **Home** tab, go to the down arrow next to the **Sort & Filter** button and select **Filter**. You will see small down arrows appear next to the heading(s) of the column(s) you had selected. These arrows represent your ability to filter the content of the column(s).

Note that if you want to change which columns have the filtering option, it is necessary to go back to the **Sort & Filter** > **Filter** option to turn off filtering for the current selection, before you can select new columns to filter.

While you are looking at the filtered view of your data, it is possible to do bulk operations (including autofill) which should not affect the data that is at that time hidden from view as it has been filtered out.

Find and Replace

Microsoft help page: Find or replace text and numbers on a worksheet

Find and **Replace** can be used to search for something in your workbook, such as a particular number or text string. They can be incredibly useful to discover what is contained in your dataset (if you want to verify how many instances of a particular value appear in the spreadsheet for example). They are also a straightforward way to implement bulk changes to a particular value, such as to correct a series of typos or spelling variations.

CTRL+F (or **Command+F** on a Mac) will allow you to run a quick **Find** search.

The full **Find** and **Replace** options can be accessed from the **Home** tab, and going to the down arrow beside the **Find & Select** button. Choosing **Replace…** will bring you to the **Find & Replace** window.

You can include wildcard characters such as question marks(?), tildes(~), and asterisks(*), or numbers in your search terms. More information can be found in the Microsoft help page linked aboved.

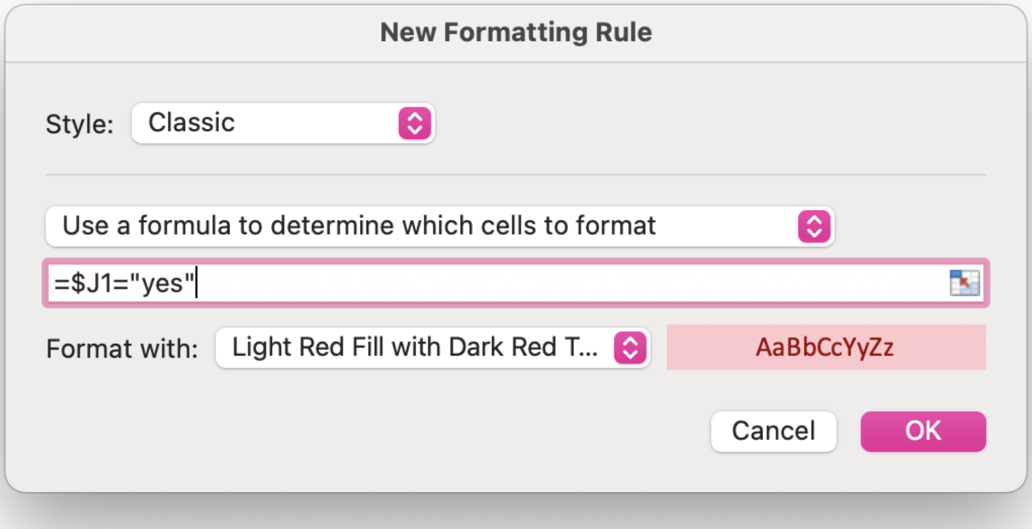
You can search by rows and columns, search within comments or values, and search the entire worksheet or workbook (see the **Options** button within the **Find & Replace** window).

Conditional Formatting

Microsoft help page: Use conditional formatting to highlight information

From Microsoft: “Conditional formatting can help make patterns and trends in your data more apparent. To use it, you create rules that determine the format of cells based on their values [...] You can apply conditional formatting to a range of cells (either a selection or a named range), an Excel table, and in Excel for Windows, even a PivotTable report.”

In our example, let’s say that we want to highlight all of the rows where the value for “agr_assoc” = “yes”. To begin, select all of the cells in this Excel sheet (you can use **CTRL+A** or **Command+A on a Mac**). On the **Home** tab, go to the drop-down arrow next to the **Conditional Formatting** button. Choose the **Manage Rules…** option. Click the **+** sign to apply a new rule. A **New Formatting Rule** window will open.



New Formatting Rule

Style: Classic

Use a formula to determine which cells to format

=$J1="yes"

Format with: Light Red Fill with Dark Red T...

Select **Classic** from the drop-down menu for **Style**. In the next drop-down, choose the option **Use a formula to determine which cells to format**. In the text box, enter the following:

=$J1=”yes”

(Because “agr_assoc” appears in column J.)

Finally, you can change the **Format** with options (to a different text/fill colour) or stick with the default. Click “OK” to apply these changes.

Conditional formatting can be removed by coming back to this same **Manage Rules…** menu.

Formulas and Functions

Mathematical Calculations

Excel is extremely powerful for executing mathematical calculations. Calculations are done through the use of formulas. For the moment, we will focus on simple mathematical formulas. Formulas can do simpler tasks, such as additions or multiplications, or more complex calculations, such as returning the cosine of an angle.

1. Formulas always begin with “=”. This tells Excel that you are not just entering numbers.
2. You can write the function yourself, or you could refer to a preset formula.
3. Math operators in Excel:
    - To add, use a plus sign: +

- To subtract, use a minus sign (hyphen): -
- To multiply, use the asterisk: *
- To divide, use the backslash: /
- Greater than is represented by the following sign: >
- Less than is represented by the following sign: <
- The math operator for “not equal” (i.e. “does not equal”) is represented by the less than and greater than signs together, like a diamond: <>

Built-In Functions

Microsoft help page: Formulas and functions

Rather than type out every calculation by hand, we can use Excel’s built-in functions. Common calculations like averages, medians, sums, and maximums have their own Excel functions.

If you think Excel may have the function you want to use you can go to the **Formulas** tab and select **Insert Function**. The functions are organized categorically in the function library (on the right-hand side). Try using the **Insert Function** button to look up how to calculate the average. Click on an empty cell, then click on **Insert Function**.

That is the first of two ways to access Excel’s built-in formulas. **Average** is under **More Functions** > **Statistical**. When you go this way, a pop-up window appears and gives you information about how to use the formula and what data the formula expects. When the formula asks for a number, you can enter a number, a cell identifier, a range, etc.

The other way is to type the name of the formula, in this case **Average**. Once you start typing “**=A**” a dropdown will appear. You can double-click on the formula you want. A shadow explanation will show up, helping you understand the formula.

Note that “usefulness” of functions will vary greatly depending on your role and the types of tasks that you are trying to complete. That being said, here are some useful functions to know:

- **IFS** Function
- Splitting and Substitution Functions
- **CONCATENATE**
- **VLOOKUP**

IFS Function

Microsoft help pages: IF function and IFS function

From Microsoft: “The **IFS** function checks whether one or more conditions are met, and returns a value that corresponds to the first **TRUE** condition. **IFS** can take the place of multiple nested **IF** statements, and is much easier to read with multiple conditions.”

(Simple) Syntax
=IFS(Something is True1, Value if True1, Something is True2, Value if True2, Something is True3, Value if True3)

Note that if no **TRUE** conditions are found, this formula will return the error message **#N/A**.

We will experiment with adding an **IFS** function to check if the following conditions are being met:

- parents_liv = yes
- sp_parents_live = yes
- grand_liv = yes
- sp_grand_liv = yes

Let’s say we want to see if any one of these conditions is true. Let’s insert a new blank column in column S to hold the results of this **IFS** query. Right-click on the S header, and go to **Insert**. Let’s name this column “IFS”. Starting in S2, we will type:

=IFS(O2="yes", "yes", P2="yes", "yes",Q2="yes", "yes",R2="yes","yes")

And then press **Enter**. You should see the word **yes** (without quotation marks) in S2.

(Note: When we are entering text in a formula, it is necessary to put it in quotation marks as we have done above, so that Excel recognizes it as a text string.)

We can use **Autofill** to carry this formula down the rest of the column.

Splitting and Substitution Functions

Microsoft help pages: Split text into different columns with functions and Split text into different columns with the Convert Text to Columns Wizard and SUBSTITUTE function

We are going to use the **Convert Text to Columns Wizard** to split the text that appears in the “liv_owned” column (column AW). To begin, we will need some additional blank columns to hold the data after the split. Highlight 10 columns to be safe, starting with the column to the right of “liv_owned” (column AX) and going to the right (“liv_owned_other”) and going to the right to “gps_Latitude” (column BG) inclusively, and then right-click and choose **Insert**. Then, click on the letter above the“liv_owned” heading to select this column. Then go to **Data > Text to Columns…** We will choose the **Delimited** option, and click **Next**. We will select our delimiter to be the **Semicolon** in this case, and then click **Next**. Finally, we will keep the data in the **General** format. Finally, you can delete any additional blank columns that are leftover.

Do you see the limitations of using this function to split data into multiple columns? We could use the **SUBSTITUTE** function to remove unwanted characters that are leftover. For example, insert a column to the left of the original “liv_owned” column, and start with this formula in the first cell:

=SUBSTITUTE(AW2,"["","")

Alternatively, a combo of **Filtering** and **Find and Replace** could be used to clean up the data at this stage. **Filtering** can also be used to identify and delete leftover blank columns from earlier in this section.

CONCATENATE

Microsoft help page: CONCATENATE function

“Use CONCATENATE, one of the text functions, to join two or more text strings into one string.”

Syntax
=CONCATENATE(text1,[text2],...)

We will use the **CONCATENATE** function to create a full version of each address (including “village”, “ward”, “district”, “province”). Start with the “village” column. We will insert a column to the right of “village”. To do this, select the “years_farm” column (column I) and right-click and choose **Insert**. We will call this new column “address_complete” (you can type this in cell I1). Then in cell I2 we will type the following formula:

=CONCATENATE(H2,", ",G2,", ",F2,", ",E2)

And then press **Enter**. You should see the address given as: **49, Bandula, Manica, Manica**.

Note that “, “ (in quotes, a comma followed by a space) is required to have a separator between each string. Otherwise, the words would be presented one after the other, with no spaces between them.

We can use **Autofill** to carry this formula down the rest of the column.

VLOOKUP

Microsoft help page: VLOOKUP function

Sorting and filtering your data can be helpful for organization. However, sometimes you just need to look up a particular value, and if you have a lot of data then sorting and filtering will still take time. The function **VLOOKUP** allows you to search your data by one variable and return the value of another variable. Another use of **VLOOKUP** is merging two datasets that have overlapping information.

Syntax
=VLOOKUP(lookup_value, table_array, col_index_num, [range_lookup])

Note: The VLOOKUP Quick Reference Card is very handy when you need a reminder of which each element of the syntax means.

Also note: **XLOOKUP** is an updated version of **VLOOKUP**, which “works in any direction and returns exact matches by default, making it easier and more convenient to use than its predecessor”. For more information, please see this function’s Microsoft help page: XLOOKUP function

We won’t go through a complete example of **VLOOKUP** together today, unless we perhaps have some extra time later. (In any case, in my experience, this is one where I need to revisit the documentation every time I use it.) But I can give an example of where it has been most useful to me:

- I had one worksheet with complete bibliographic information for an ebook collection. Their MARC records had been mapped to columns in Excel, including title, author, OCLC number (which is a persistent and unique identifier), current URL, coverage information, etc.
- In a second worksheet, I had the OCLC numbers listed alongside new URLs which had been provided by the vendor because the URL structure was changing.
- In order to bulk update the links for this collection for the McGill catalog, I used **VLOOKUP** to connect the new URLs to the appropriate bibliographic record via their OCLC numbers. Then I could use the newly linked information together to create a KBART file for upload to the McGill knowledge base and catalog.

Charts

Excel offers the opportunity to do data visualizations via the **Chart** functionality. We will not go into detail about **Charts** during today’s workshop, but if you ever need to create a **Chart**, know that this functionality exists. Microsoft has some helpful documentation that can guide you through creating a chart: Create a chart from start to finish

PivotTables

Microsoft help page: Overview of PivotTables and PivotCharts

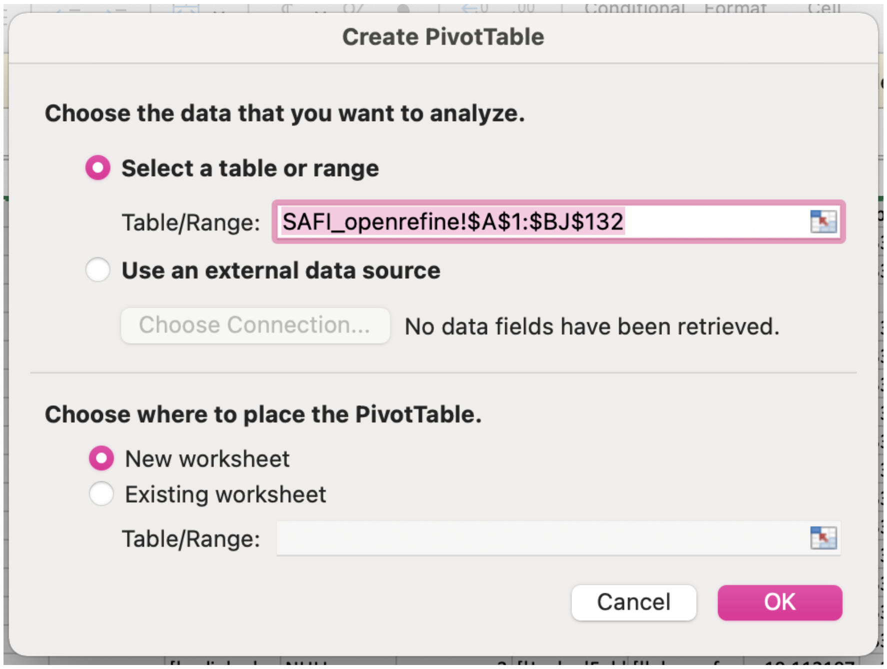
From Microsoft: “You can use a PivotTable to summarize, analyze, explore, and present summary data. PivotCharts complement PivotTables by adding visualizations to the summary data in a PivotTable, and allow you to easily see comparisons, patterns, and trends.”

[...]

“A PivotTable is an interactive way to quickly summarize large amounts of data. You can use a PivotTable to analyze numerical data in detail, and answer unanticipated questions about your data. A PivotTable is especially designed for:

- Querying large amounts of data in many user-friendly ways
- Subtotally and aggregating numeric data, summarizing data by categories and subcategories, and creating custom calculations and formulas
- Expanding and collapsing levels of data to focus your results, and drilling down to details from the summary data for areas of interest to you
- Moving rows to columns or columns to rows (or ‘pivoting’) to see different summaries of the source data
- Filtering, sorting, grouping, and conditionally formatting the most useful and interesting subset of data enabling you to focus on just the information you want
- Presenting concise, attractive, and annotated online or printed reports”

To begin, select all the data on your worksheet using **CTRL+A** (or **Command+A** on a Mac). Then go to **Data** > **Summarize with PivotTable**, or go to the **Insert** tab and click the **PivotTable** button. This will open the **Create PivotTable** menu.



Create PivotTable

Choose the data that you want to analyze.

● Select a table or range

Table/Range: SAFI_openrefine!$A$1:$BJ$132

○ Use an external data source

Choose Connection... No data fields have been retrieved.

Choose where to place the PivotTable.

● New worksheet

○ Existing worksheet

Table/Range:

Cancel OK

Verify that all of the data has been selected in the **Table/Range** field. For **Choose where to place the PivotTable** is usually easiest to choose **New worksheet**. It is not necessary to specify a **Table/Change**. Then press **OK**.

Your PivotTable interface should load in a new sheet (Sheet1). Feel free to rename the sheet at this stage, by right-clicking on the tab that says Sheet1 and choosing **Rename**. You can also drag the tab for this sheet to the right-hand side of the data tab, if you prefer (I usually do this as it feels more ergonomic to me).

You will see a variety of options in the **PivotTable Fields** menu on the right-hand side of the page. If you ever accidentally close this menu, simply right-click on the PivotTable1 placeholder:

PivotTable1

Click in this area to work with the PivotTable report

or your active PivotTable and choose **Show Field List**. Note that clicking away from the PivotTable1 placeholder or the active PivotTable will also cause the **PivotTable Fields** menu to be hidden, but clicking on it again will ensure it shows again.

The **PivotTable Fields** menu has 5 important components:

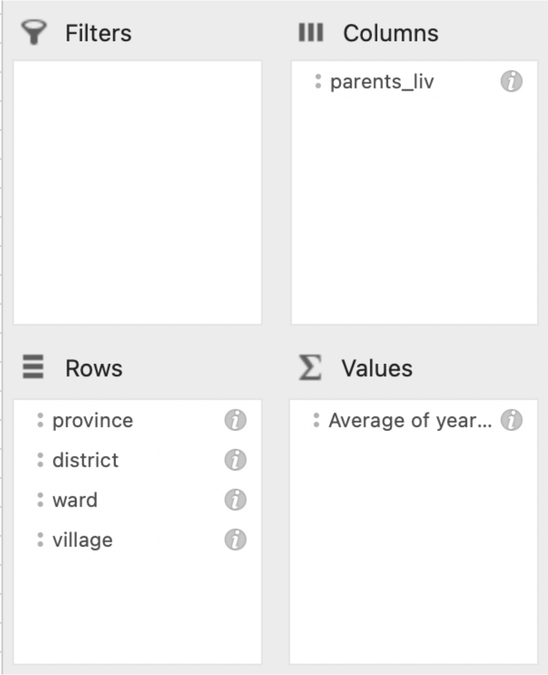
- **Field Name**
    - Contains all of the variables names, i.e. all of the header names of the columns in your dataset
- **Filters**
- **Columns**
- **Rows**
- **Values**

We will experiment moving variables from **Field Name** among each of the other four areas to build a PivotTable. This goal is to discover trends and patterns in our data that we might not otherwise see.

A few things to note:

- To include a field name, check the box beside the name within the **Field Name** box.
- The field name will appear in one of the four boxes below. It can be dragged and dropped within each category.
- Order within each category (**Filters**, **Columns**, **Rows**, **Values**) matters. You can also drag and drop the order of field names within a given category.
- Clicking on the little ***i*** to the right of each field name allows you to customize each field, such as changing from **count** to **average.**

Feel free to experiment! For today, I will be working towards the following:



Filters

Columns
- parents_liv

Rows
- province
- district
- ward
- village

Values
- Average of year...

This PivotTable utilizes the following variables in the **Field Name** box:

- province
- district
- ward
- village
- years_liv
- parents_liv

In **Rows** you will have (in this exact order):

- province
- district
- ward
- village

Under **Columns** you will have:

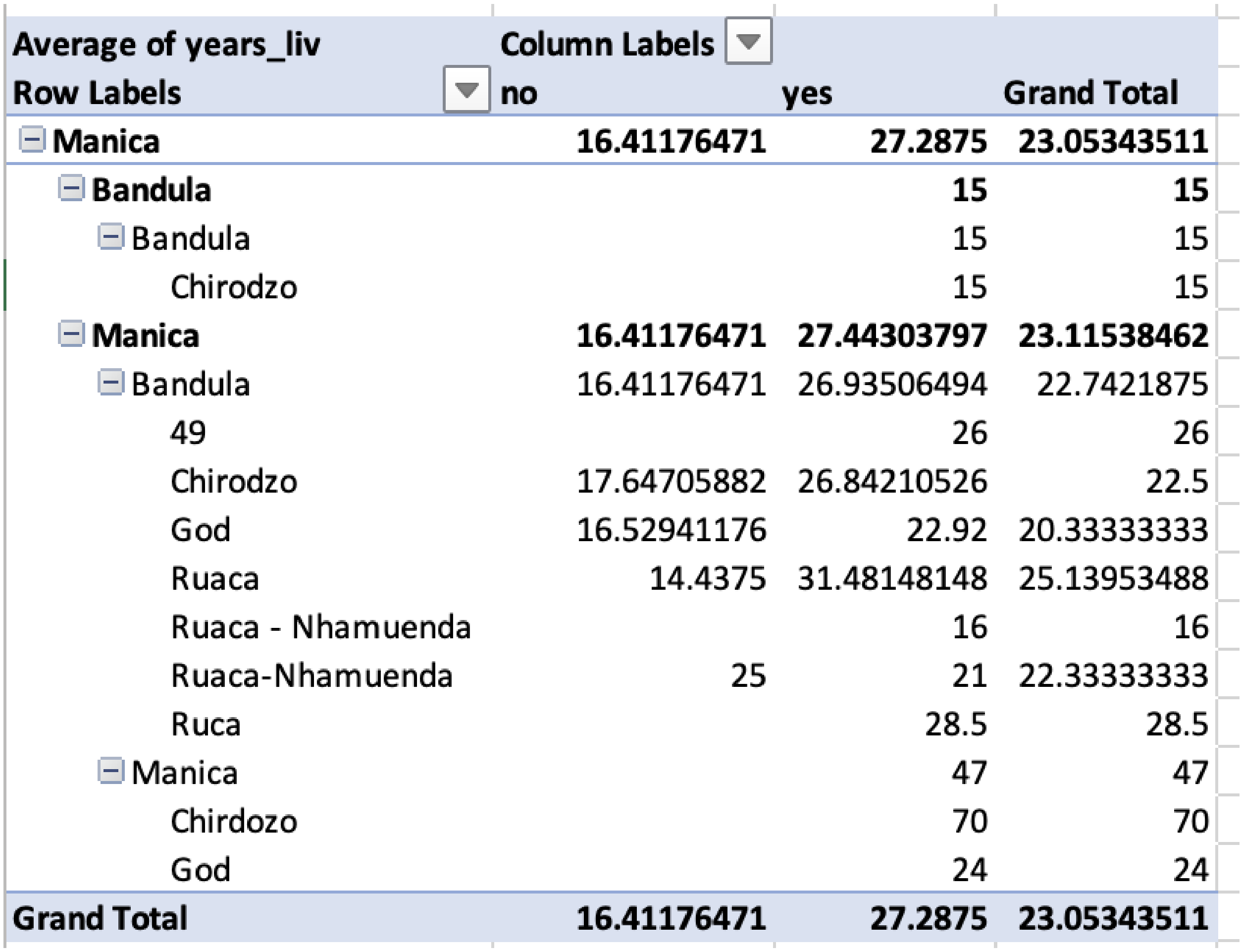
- parents_liv

And finally, under **Values** you will have:

- year_liv

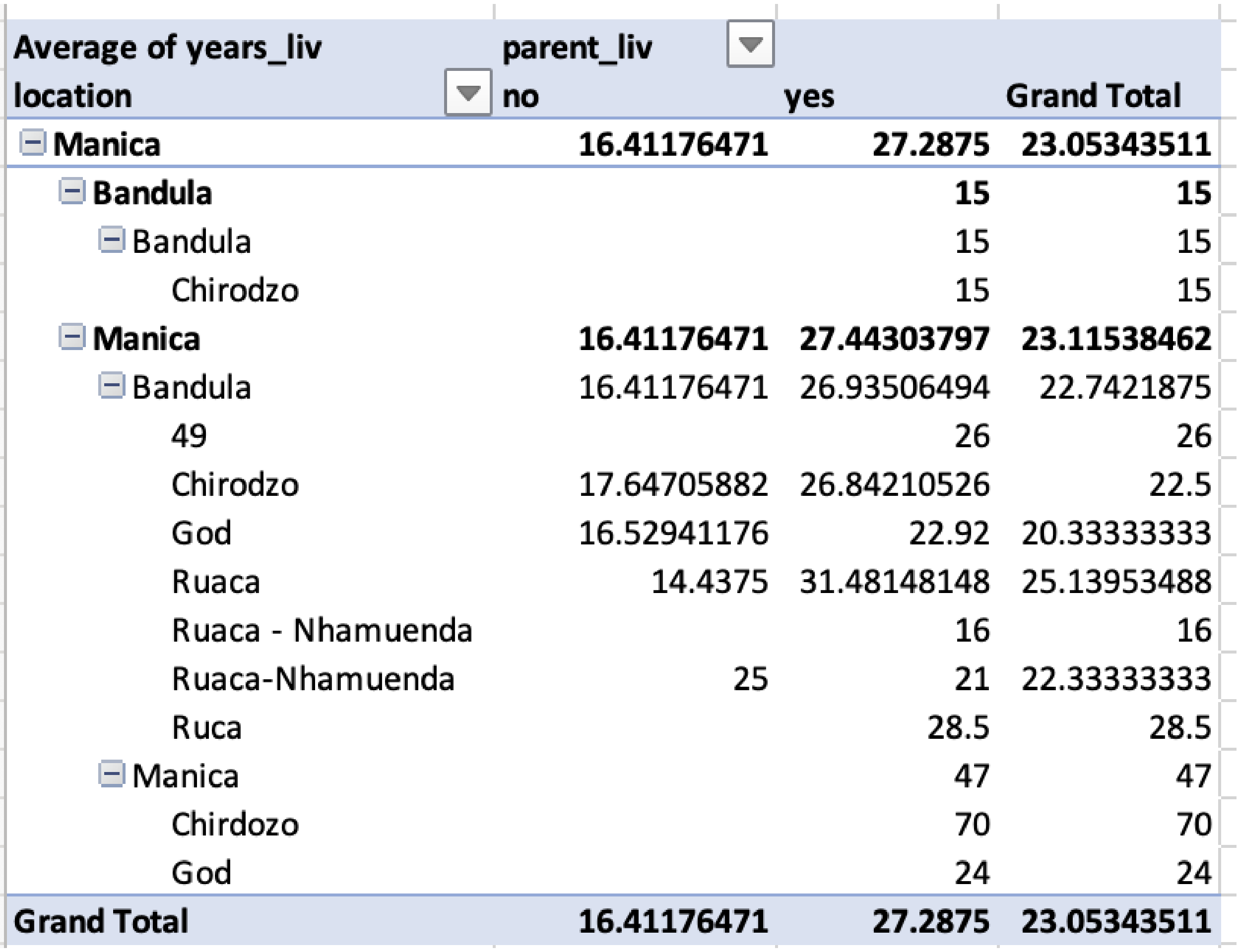
And you will need to go to the *i* symbol beside year_liv to change the **Summarize by** option to be **Average**.

Which will create a PivotTable that looks like this:



Average of years_liv	Column Labels ▼		
Row Labels ▼	no	yes	Grand Total
⊟ **Manica**	**16.41176471**	**27.2875**	**23.05343511**
⊟ **Bandula**		**15**	**15**
⊟ Bandula		15	15
Chirodzo		15	15
⊟ **Manica**	**16.41176471**	**27.44303797**	**23.11538462**
⊟ Bandula	16.41176471	26.93506494	22.7421875
49		26	26
Chirodzo	17.64705882	26.84210526	22.5
God	16.52941176	22.92	20.33333333
Ruaca	14.4375	31.48148148	25.13953488
Ruaca - Nhamuenda		16	16
Ruaca-Nhamuenda	25	21	22.33333333
Ruca		28.5	28.5
⊟ Manica		47	47
Chirdozo		70	70
God		24	24
Grand Total	**16.41176471**	**27.2875**	**23.05343511**

As a final cleanup step, the cells that say **Column Labels** and **Row Labels** can be renamed simply by clicking and typing in these cells. For example, I might change **Column Labels** to **parent_liv** and **Row Labels** to **location**, like so:



Average of years_liv	parent_liv ▼		
location ▼	no	yes	Grand Total
⊟ Manica	**16.41176471**	**27.2875**	**23.05343511**
⊟ Bandula		**15**	**15**
⊟ Bandula		15	15
Chirodzo		15	15
⊟ Manica	**16.41176471**	**27.44303797**	**23.11538462**
⊟ Bandula	16.41176471	26.93506494	22.7421875
49		26	26
Chirodzo	17.64705882	26.84210526	22.5
God	16.52941176	22.92	20.33333333
Ruaca	14.4375	31.48148148	25.13953488
Ruaca - Nhamuenda		16	16
Ruaca-Nhamuenda	25	21	22.33333333
Ruca		28.5	28.5
⊟ Manica		47	47
Chirdozo		70	70
God		24	24
Grand Total	**16.41176471**	**27.2875**	**23.05343511**